INSTITUTO
UNIVERSITÁRIO
DE LISBOA

From Regular Expressions to Language Models: A Text-Based Approach for Invoice Information Extraction

Marta Gameiro Francisco Moreno Pereira

Master in Integrated Business Intelligence Systems

Supervisor:
PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – IUL

Co-Supervisor:
PhD Fernando Manuel Marques Batista, Associate Professor, Iscte – IUL

October, 2024

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

From Regular Expressions to Language Models: A Text-Based Approach for Invoice Information Extraction

Marta Gameiro Francisco Moreno Pereira

Master in Integrated Business Intelligence Systems

Supervisor:
PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – IUL

Co-Supervisor:
PhD Fernando Manuel Marques Batista, Associate Professor, Iscte – IUL

October, 2024

Acknowledgment

I would like to express my deepest gratitude to everyone who has supported me throughout this journey.

First and foremost, I am incredibly grateful to my professors, Ricardo Ribeiro and Fernando Batista, whose guidance and expertise were invaluable to this dissertation and the work it represents. Their constant availability, encouragement, and commitment to excellence pushed me to strive harder and achieve more than I could have anticipated. This work would not have been possible without their dedication and support.

To my girlfriend, I am deeply appreciative of her belief in me, even through challenging times. Her unwavering faith in my abilities and steadfast support provided me with the strength to persevere, and I am truly fortunate to have had her by my side.

To my friends, I am thankful for their understanding and support. The support and encouragement they provided were incredibly valuable, especially when I couldn't spend as much time with them. I am grateful to each of them for standing by me throughout this process.

Finally, I want to extend my heartfelt gratitude to my family, especially my mother, whose encouragement inspired me to pursue my studies and aim higher. Her belief in my potential and her constant motivation have been my greatest source of strength. I am profoundly appreciative of the support my family has provided, which has been essential to completing this work.

To each of you, thank you.

Resumo

A automatização da extração de informações de faturas através de aprendizagem automática apresenta benefícios significativos para as organizações, especialmente na melhoria da eficiência e precisão. A implementação de algoritmos de aprendizagem automática permite às empresas reduzir o tempo e o esforço dedicados à introdução e validação manual de dados, permitindo que as empresas se concentrem em atividades mais estratégicas. Contudo, empresas globais enfrentam desafios particulares ao lidar com faturas em diversas línguas e formatos distintos, o que exige soluções personalizadas.

Apesar do crescente interesse nesta temática, são escassos os estudos que abordam técnicas baseadas em texto para a extração de informação de faturas. Esta lacuna na literatura evidencia a necessidade de mais investigações para otimizar metodologias que visam extrair dados como o NIF, datas de emissão e valores totais.

O presente trabalho, baseado num conjunto de dados fornecido por uma empresa privada, permite realizar uma análise em contexto real. O objetivo foi desenvolver técnicas eficazes de extração e avaliar a sua aplicabilidade no ambiente da empresa. Foram realizados dois testes: expressões regulares alcançaram 75% de precisão na extração do NIF, enquanto modelos generativos, especificamente o Mistral-7B-Instruct-v0.3, obtiveram 76% para datas e 71% para valores totais. Estes resultados indicam que, aprofundando o estudo, testando técnicas adicionais e superando algumas limitações, especialmente no que diz respeito a restrições de privacidade, as metodologias de extração podem ser refinadas, evidenciando um grande potencial para a automação de extração de informações de faturas.

Palavras Chave: *Extração de Informação*, *Processamento de Faturas*, *Aprendizagem Automática*, *Expressões Regulares*, *Modelos de Linguagem*, *Automação*.

Abstract

Automating the extraction of invoice information through Machine Learning (ML) offers significant benefits to organizations, particularly in enhancing efficiency and accuracy. By leveraging ML algorithms, businesses can significantly reduce the time and effort required for manual data entry and validation, allowing them to focus on more strategic activities. However, global companies face unique challenges when handling multilingual invoices and documents with varying layouts, which complicates the extraction process and necessitates the development of tailored solutions.

Despite the growing interest in this topic, there are relatively few studies focusing on text-based techniques for information extraction from invoices. This gap in the literature highlights the need for further research to identify and optimize effective methodologies for extracting relevant data fields such as Value Added Tax (VAT) numbers, issuance dates, and total values.

The present work, based on a dataset provided by a private company, allows for an analysis in a real-world context. The goal was to develop effective extraction techniques and assess their applicability in the company's environment. Two tests were conducted: regular expressions achieved 75% accuracy for VAT extraction, while generative models, specifically Mistral-7B-Instruct-v0.3, achieved 76% accuracy for date and 71% for total value extraction. This performance suggests that, by deepening the study, testing additional techniques, and overcoming limitations, particularly with respect to privacy restrictions, extraction methodologies could be further refined, revealing substantial potential for automating invoice information extraction.

KEYWORDS: *Information Extraction*, *Invoice Processing*, *Machine Learning*, *Regular Expressions*, *Language Models*, *Automation*.

Contents

Acknowledgment i

Resumo iii

Abstract v

List of Figures ix

List of Tables xi

List of Acronyms xiii

Chapter 1. Introduction 1
1.1. Motivation 2
1.2. Research Questions and Goals 4
1.3. Structure and Organization of the Dissertation 6

Chapter 2. Related Work 7
2.1. Method 7
2.2. Data Extraction 7
2.3. State of the Art 9
2.3.1. One-dimensional Tools 11
2.3.2. Two-dimensional Tools 12
2.3.3. Large Language Models 17

Chapter 3. Materials and Methods 19
3.1. Data 19
3.2. Methods 21
3.2.1. PDF to Text 23
3.2.2. Knowledge Extraction 24

Chapter 4. Experiments and Results 29
4.1. Regular Expressions 30
4.1.1. Baseline 30
4.1.2. Advanced Pattern Extraction 31
4.2. LLMs and Generative Models: Contextual and Semantic Understanding 36
4.3. Discussion 40

Chapter 5. Conclusions and Future Work 43

5.1. Conclusions 43
5.2. Limitations and Future Work 45

References 47

List of Figures

Figure 1.1 A Review From a Resource-oriented Perspective on How Big Data Influences Competitive Advantages. 2

Figure 1.2 An Example Invoice. 5

Figure 2.1 Documents by Year. 9

Figure 2.2 Documents by Subject Area. 9

Figure 2.3 Documents by Keyword. 10

Figure 2.4 Four stages of audit workflow. 12

Figure 2.5 Workflow of the Chargrid method. The document is encoded as a 2D grid of characters, preserving spatial structure. The process includes semantic segmentation, pixel-level labeling, and object-bounding box predictions for information extraction. 13

Figure 2.6 Workflow for the approach by [24] . The method integrates pre-trained language models with Graph Neural Networks (GNN) to encode textual and visual features from digital documents while preserving rich formatting and metadata. 14

Figure 3.1 Invoice Example From the Portuguese Invoices Training Set. 22

Figure 3.2 Processing Pipeline. 23

Figure 3.3 Original invoice image 26

Figure 3.4 Text extracted from the invoice using Tesseract OCR 27

Figure 4.1 Baseline Regular Expressions for Each Field. 31

Figure 4.2 Complex Regular Expression for VAT Number Field. 32

Figure 4.3 Complex Regular Expression for Date Field. 34

Figure 4.4 Complex Regular Expression for Total Value Field. 35

Figure 4.5 Baseline Prompt. 38

Figure 4.6 Detailed Prompt. 39

List of Tables

Table 2.1	Definition of Keywords Used for Searching the Scopus Database.	8
Table 3.1	Simplified Structure Defined by the Portuguese Tax Authority for Generating QR Codes.	20
Table 4.1	Extraction Accuracy of VAT Number, Date, and Total Value Fields Using Different Techniques.	40

List of Acronyms

BERT: Bidirectional Encoder Representations from Transformers

CFG: Context-free Grammar

DeepCPCFG: Deep Conditional Probabilistic Context-Free Grammars

DL: Deep Learning

EIN: Employer Identification Number

GNN: Graph Neural Networks

GPT: Generative Pre-trained Transformer

IBAN: International Bank Account Number

IE: Information Extraction

ITE: Image Text Extraction

JPEG: Joint Photographic Experts Group

KIE: Key Information Extraction

LLM: Large Language Model

LLMs: Large Language Models

MC-OCR: Mobile-Captured Image Document Recognition

ML: Machine Learning

NaN: Not a Number

NER: Named Entity Recognition

NIF: Número de Identificação Fiscal

NIPC: Número de Identificação de Pessoa Coletiva

NLP: Natural Language Processing

OCR: Optical Character Recognition

PSM: Page Segmentation Mode

PDF: Portable Document Format

PNG: Portable Network Graphics

QR Code: Quick Response Code

QR Codes: Quick Response Codes

RIVF: Research, Innovation and Vision for the Future

RNPC: Registo Nacional de Pessoas Coletivas

SPADE: SPAtial DEpendency parser

VAT: Value Added Tax

VRD: Visually Rich Documents

CHAPTER 1

Introduction

In today's digital age, businesses generate a massive amount of data on a daily basis. According to Gärtner and Hiebl [1], for management accountants, unstructured data represents 80% of all available data. From Natural Language Processing (NLP) to Machine Learning (ML) algorithms, unstructured data has become a cornerstone in driving innovation and efficiency across multiple domains. One such domain where unstructured data plays a pivotal role is the extraction of invoice information.

Managing and analyzing these documents is a costly and time-consuming process for organizations. In the accounting field, tasks such as invoice information extraction often require a significant amount of time and are mainly conducted manually. The process involves meticulously reviewing each invoice and extracting relevant information such as invoice number, date, and totals. This manual process can be prone to errors and is not scalable for businesses with a large volume of invoices. As a result, there is a growing need for automated solutions to streamline the extraction of invoice information and improve the efficiency of accounting processes.

In several regions, including Portugal and Poland, regulatory measures have mandated the inclusion of QR codes on invoices. These QR codes contain essential information, offering an alternative means for data extraction and serving as a method for validating automated systems. However, the adoption of QR codes in invoices has yet to become a universal standard, which presents challenges for the broader applicability of this approach.

Moreover, the use of QR codes is not without limitations. Problems such as misreading the codes or difficulties in establishing connections to servers during data retrieval can introduce uncertainties regarding the accuracy and reliability of the extracted information. These challenges underscore the importance of developing robust, text-based techniques that can either complement or serve as substitutes for QR code-based methods when necessary.

This study investigates the importance of extracting valuable data from invoices, revealing key findings that are essential for improving financial operations, making better decisions, and optimizing workflow efficiency. A rise in applications utilizing unstructured data mining is being observed in the accounting and management sectors. These tools have the potential to facilitate real-time audits, prevent fraud, and improve forecasting accuracy [2].

Figure 1.1 by Erevelles, Fukawa, and Swayne [3], demonstrates that dynamic and adaptive capabilities, obtained from analyzing structured, semi-structured, and unstructured

data, play a crucial role in creating value. Big Data enables improved decision-making which leads to the creation of either sustainable or temporary competitive advantages.

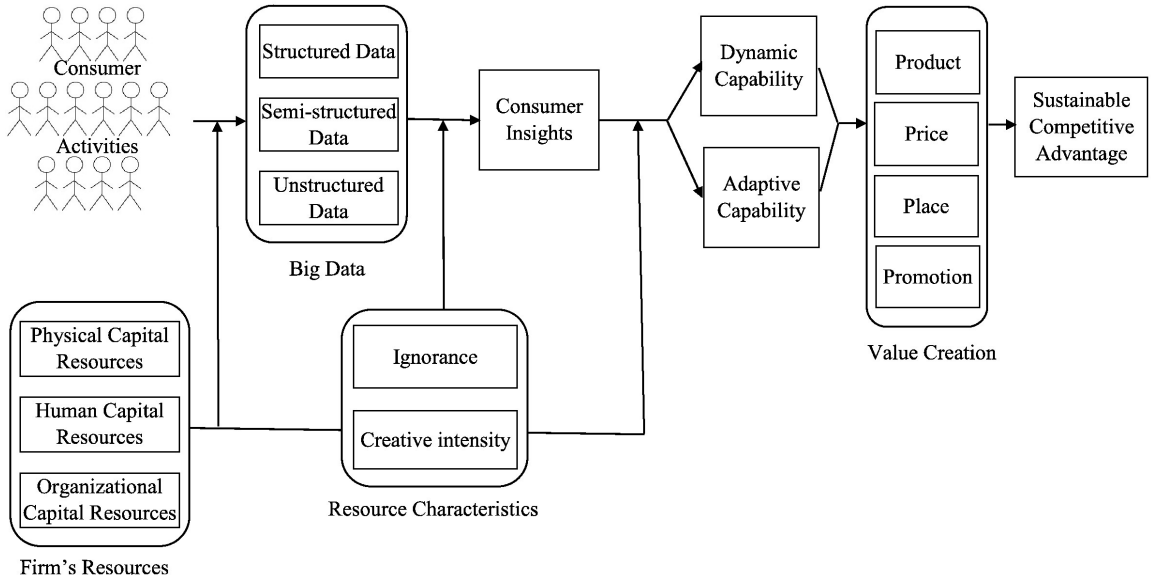


FIGURE 1.1. A Review From a Resource-oriented Perspective on How Big Data Influences Competitive Advantages.

1.1. Motivation

The increasing volume of financial documents has heightened the importance of efficient and accurate invoice processing. Manual extraction of information is time-consuming and error-prone, leading to workflow bottlenecks and inaccuracies in financial reporting. To address these challenges while aiming for scalability and precision, the automation of invoice data extraction is essential.

One promising development in this field is the integration of Quick Response Codes (QR Codes) into invoices, a requirement in certain regions such as Portugal and Poland. QR Codes provide an efficient way to access essential invoice data in a structured and machine-readable format. By encoding vital information, including invoice numbers, dates, totals, and tax details, QR Codes can significantly simplify the processes of data extraction and validation. Their adoption serves as a valuable resource for ensuring data accuracy and enhancing processing speeds, thereby addressing some of the challenges posed by manual workflows.

Despite their advantages, QR Codes have limitations. Their adoption is inconsistent worldwide, with many countries yet to establish regulations mandating their use. This inconsistency presents challenges for international organizations, as not all invoices incorporate this feature, leading to variability in processing. Additionally, the reliability of QR Codes can be compromised by technical issues such as poor printing quality, physical damage, or errors during code generation. In some instances, incomplete or incorrect information within the QR Codes can further complicate data extraction, necessitating supplementary validation methods.

These challenges underscore the need to consider QR Codes as a complementary rather than a standalone solution for invoice processing. They highlight the broader necessity for robust and adaptable approaches that can accommodate various document formats, address technical challenges, and function effectively even in the absence of QR Codes. This reality serves as a crucial impetus for developing more comprehensive techniques, ensuring that organizations can reliably and efficiently extract the data required to enhance their operations.

In light of these challenges, automating the extraction of invoice information through ML offers significant benefits to organizations. By leveraging ML algorithms, businesses can streamline their processes, reducing the time and effort required for manual data entry and validation. This not only improves the accuracy of extracted information but also frees up valuable resources to focus on more strategic tasks [4]. Additionally, implementing automated systems in invoice processing allows organizations to handle a larger volume of invoices without the need for additional manual intervention, leading to improved scalability and efficiency.

Global companies face unique challenges when it comes to handling multi-language invoices and documents with different layouts. Invoices from different countries can be written in different languages, making it difficult to extract information using traditional methods. ML models used for Information Extraction (IE) need to be robust enough to handle and interpret various languages and character sets.

Additionally, invoices from different regions may have different layouts, formats, and even currencies. This adds another layer of complexity to the extraction process. ML models must be trained on a diverse set of templates and formats to accurately extract information regardless of the variations in the layout.

Deep Learning approaches can continuously learn and adapt to new formats and variations of invoices, making them highly adaptable to changing business needs. This adaptability ensures that the extraction process remains robust even as the volume and complexity of invoices evolve over time.

The use of these architectures in invoice processing also aligns with the broader trend of digital transformation in business operations. It enables organizations to streamline their accounting workflows, improve data accuracy, reduce manual errors, and gain deeper insights from the extracted invoice data. However, as the landscape of business documents becomes more complex, organizations face significant challenges in handling diverse and intricate unstructured documents [5]. This complexity underscores the need for more sophisticated tools that can efficiently process a wide range of invoice formats and structures.

The motivation for developing an automated tool for invoice processing stems from the potential to enhance operational efficiency, reduce costs, and support more informed decision-making through access to accurate and timely invoice data. Traditional manual processing methods can be labor-intensive, time-consuming, and susceptible to human

errors, all of which can lead to delays in financial operations and increased costs. An automated approach offers the opportunity to overcome these issues by enabling faster processing times, minimizing the potential for errors, and freeing up human resources for more strategic tasks.

The majority of research and tools in automated invoice processing have concentrated on multi-dimensional approaches, integrating both textual and structural analysis. However, there is a notable lack of studies on one-dimensional tools that focus solely on text-based analysis. Despite the rise of advanced, multi-dimensional solutions, there is a gap in the exploration of simpler, one-dimensional tools that could provide more lightweight and accessible solutions. These tools, by concentrating exclusively on text extraction and analysis, could offer a practical alternative in scenarios where a full structural understanding of the document is not required or where computational resources are limited.

This gap highlights an opportunity to develop a one-dimensional tool focusing on text-based analysis for invoice processing. Such an approach has the potential to provide a streamlined, resource-efficient solution that meets the needs of organizations seeking simpler, more cost-effective automation options. By addressing this overlooked area, the proposed work aims to offer a viable alternative to more complex models, ensuring that even smaller businesses or those with limited technical resources can benefit from automated invoice processing. Ultimately, this tool could enhance data accessibility, improve operational efficiency, and support organizations in making timely and accurate business decisions.

1.2. Research Questions and Goals

The primary objective of this research is to construct an architecture dedicated to the extraction of relevant information from invoices, as illustrated in figure 1.2 from Saout, Lardeux, and Saubion [6]. The resulting tool, designed through this study, aims to seamlessly process invoices, extract key information such as invoice numbers, dates, and total amounts, and ultimately organize these data into a format conducive to computer readability.

To achieve this, the solution employs advanced data extraction and processing methodologies, leveraging ML algorithms, along with NLP techniques. The focus is on building a system that not only retrieves relevant data but also ensures that the data is structured coherently, enabling seamless integration into existing computer systems and databases.

A fundamental aspect of this work is the need to address the challenges posed by the diverse landscape of invoices. Invoices often differ significantly in terms of layout, formatting, and linguistic content, which adds complexity to the extraction process. A notable challenge is the variability in the languages present in invoices, which requires the creation of a solution capable of recognizing and handling information across multiple languages, each with its own syntax, structure, and linguistic nuances. The ability to process multilingual invoices is crucial for ensuring that the developed tool can adapt to various international contexts, making it a versatile solution in global applications.

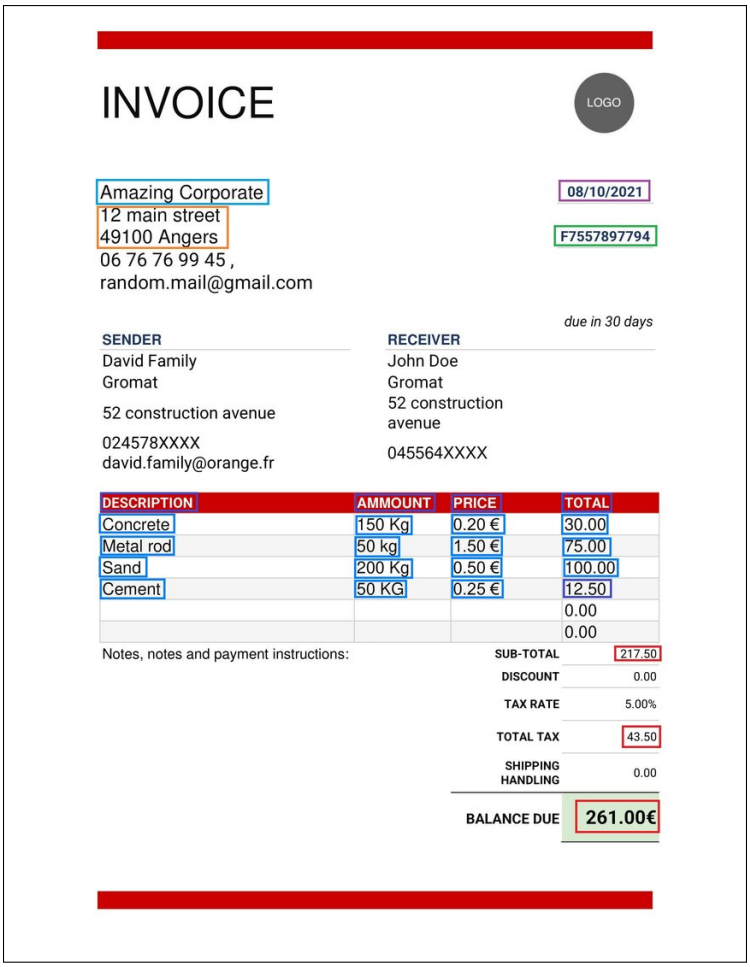


FIGURE 1.2. An Example Invoice. Fonte: Saout, Lardeux, and Saubion [6]

This study is guided by two central research questions:

(1) How can text-based information extraction methods be optimized to accurately extract relevant fields (e.g., invoice number, date, total amount) from multilingual invoices?
(2) How effective are pre-trained multilingual models for information extraction tasks on invoice text data?

These questions are pivotal to understanding how the proposed architecture can be refined to address the linguistic diversity and variability in invoice structures. By examining the optimization of extraction techniques, the study seeks to enhance the precision and accuracy of the tool, ensuring that it can identify key data fields even in challenging scenarios.

Furthermore, exploring the efficacy of pre-trained multilingual models directly informs the development of the solution by assessing how well these models perform in real-world settings. This investigation aims to determine whether existing models can be adapted to the specific needs of invoice processing or if custom models are required to achieve the desired accuracy. This exploration endeavors not only to advance the boundaries of

current information extraction methodologies but also to enrich the broader academic comprehension of applications in multilingual natural language processing.

By addressing these research questions and employing a combination of ML, Deep Learning (DL), and NLP techniques, the study aims to develop a robust, adaptable solution for information extraction from invoices. This solution aspires to generalize its understanding to accommodate a wide array of layout variations and adapt to different languages, emerging as a comprehensive tool for automated invoice processing.

1.3. Structure and Organization of the Dissertation

This dissertation is organized into six chapters, each addressing a critical aspect of the research process.

Chapter 1: Introduction provides an overview of the motivation behind this study, outlines the research questions and goals, and presents the structure of the document itself. It establishes the context for the problem of information extraction from invoices and highlights the significance of the study.

Chapter 2: Related Work reviews existing literature and methodologies in the domain of data extraction, focusing on both traditional and advanced approaches. It explores one-dimensional tools, such as regular expressions, as well as two-dimensional tools that leverage complex spatial relationships within document layouts. This chapter situates the research within the broader academic context.

Chapter 3: Data details the dataset used for this research, including the sources, preprocessing steps, and characteristics of the invoice data.

Chapter 4: Methods describes the techniques and algorithms employed in the development of the information extraction tool, including methods of ML, and NLP methods.

Chapter 5: Experiments and Results presents the experimental setup, evaluates various approaches, and discusses the outcomes. It covers baseline methods, such as regular expressions, as well as more advanced techniques using Large Language Models (LLMs).

Chapter 6: Conclusions and Future Work summarizes the main findings, discusses their implications, and suggests directions for future research in the field of automated information extraction.

CHAPTER 2

Related Work

The next chapter of this dissertation aims to offer a comprehensive summary of the research carried out in the area of extracting invoice information, with the aim of improving the understanding of the existing work and the advances in this field. The literature review on invoice IE will provide an overview of the research conducted in this area, highlighting the current state of the art and advances. It will also identify the challenges and limitations that researchers have encountered in this field, as well as the strategies and methodologies used to overcome them.

2.1. Method

The systematic research methodology was based on PRISMA [7], which stands for Preferred Reporting Items for Systematic Reviews and Meta-Analyses, guidelines. This methodology was adopted to ensure a rigorous and transparent approach to the literature review process. A single query was used to retrieve relevant publications from Scopus. The inclusion and exclusion criteria were defined to enhance the search results, ensuring relevance to the topic. This chapter will discuss these criteria in section 2.2, 'Data Extraction'.

2.2. Data Extraction

The data collection process involved searching the Scopus database using a single query. The keyword search was structured into three categories: Population, Concept, and Context. A detailed summary of the search strategy used, along with specific keywords for each category, can be found in Table 2.1.

Narrowing the search to the final 46 documents involved a systematic and iterative approach to ensure relevance and specificity. First, a broad search was conducted to capture documents related to the key technological concepts, including “Text Mining”, “NLP”, “Natural Language Processing”, “Machine Learning”, “Deep Learning”, and “Data Mining.” These terms were connected using an OR operator, resulting in a total of 1,142,026 documents.

A focused search was carried out to address the specific problem domain by incorporating the terms "invoice*" or "receipt*," yielding a total of 45,196 documents. By combining these conceptual keywords with problem-specific terms using an AND operator, the results were narrowed down to 560 documents. This indicates a distinct subset of research that explicitly intersects the two domains.

To incorporate the contextual dimension, a separate search was performed using the term "Information Extraction," which yielded 22,808 documents. By combining the conceptual keywords, problem keywords, and contextual keywords, the Scopus database identified 80 documents that specifically examined the intersection of these three dimensions. Additional restrictions were applied, including a limitation on publication years to 2018–2023, a focus on specific document types such as conference papers and articles, and a requirement for the language to be English. These stringent criteria resulted in a final dataset of 46 documents, which reflects a highly targeted selection of research related to the defined concepts, problems, and context.

TABLE 2.1. Definition of Keywords Used for Searching the Scopus Database.

<table>
<tr><th>Concept</th><th>Problem</th><th>Context</th><th>Limitations</th></tr>
<tr><td>Text Mining
NLP
Natural Language Processing
Machine Learning
Deep Learning
Data Mining</td><td>Invoice*
Receipt*</td><td>Information Extraction</td><td rowspan="4">Only conference papers and articles from 2018 to 2023, written in English</td></tr>
<tr><td>1 142 026 documents</td><td>45 196 documents</td><td rowspan="2">22 808 documents</td></tr>
<tr><td colspan="2">560 documents</td></tr>
<tr><td colspan="3">80 documents</td></tr>
<tr><td colspan="4">46 documents</td></tr>
</table>

In addition to the search criteria mentioned earlier, additional documents were excluded based on specific criteria. Nine documents were deemed out of scope, as they pertained to tasks unrelated to the extraction of information from invoices. Four documents were excluded because they focused on other types of business documents rather than invoices. Another four documents were removed due to redundancy, as they were variations of other papers already included in the selection. Finally, one document was excluded for being centered on a dataset, rather than addressing the task of information extraction from invoices. These exclusions were made to ensure the final corpus was strictly aligned with the scope of the study.

Figure 2.1 shows the distribution of publications over the years, demonstrating the increasing interest in this subject. In 2023, the search query yielded the highest number of publications, accounting for 26% of all results. Based on these data, it is clear that there has been a noticeable growth in attention to this topic.

The distribution of documents by area is diverse. The most common areas include Computer Science, Mathematics, and Engineering, as demonstrated by Figure 2.2. The field of computer science is present in 47% of publications, followed by mathematics at 18% and engineering at 14%.

Keyword analysis also revealed the most common keywords resulting from the search query, as shown in figure 2.3: Information Retrieval, Information Extraction, DL, Optical Character Recognition (OCR), Natural Language Processing Systems, and Data Mining. All of these keywords were referenced in more than 10 publications.

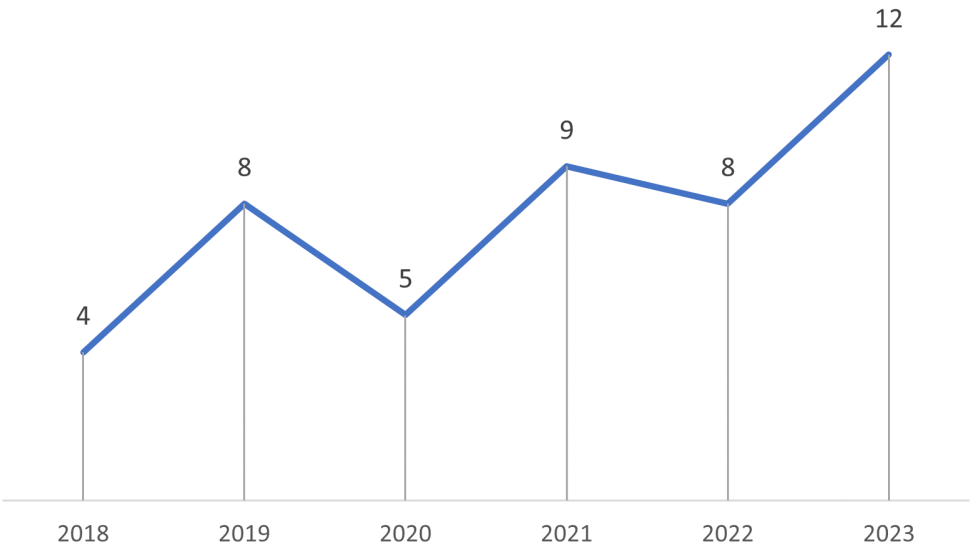


FIGURE 2.1. Documents by Year.

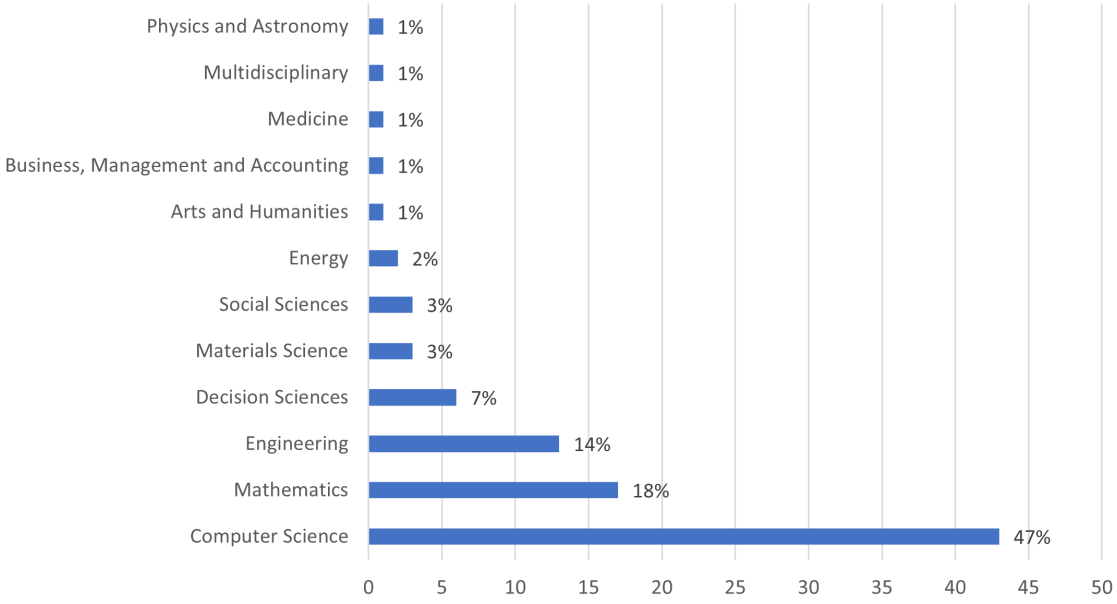


FIGURE 2.2. Documents by Subject Area.

2.3. State of the Art

Invoice IE is a critical task in automating business processes and optimizing efficiency, especially in accounting fields. Forbes [8] states that unstructured data represents 80 to 90 percent of all data in businesses, while Deloitte Insights [9] indicates that only 18% of companies take advantage of this data. This under-utilization of unstructured data, including invoice information, highlights the need for efficient extraction methods. Studies in this field are divided into two broad categories: feature extraction techniques, such as rule-based or template-based approaches, and ML approaches. Feature extraction

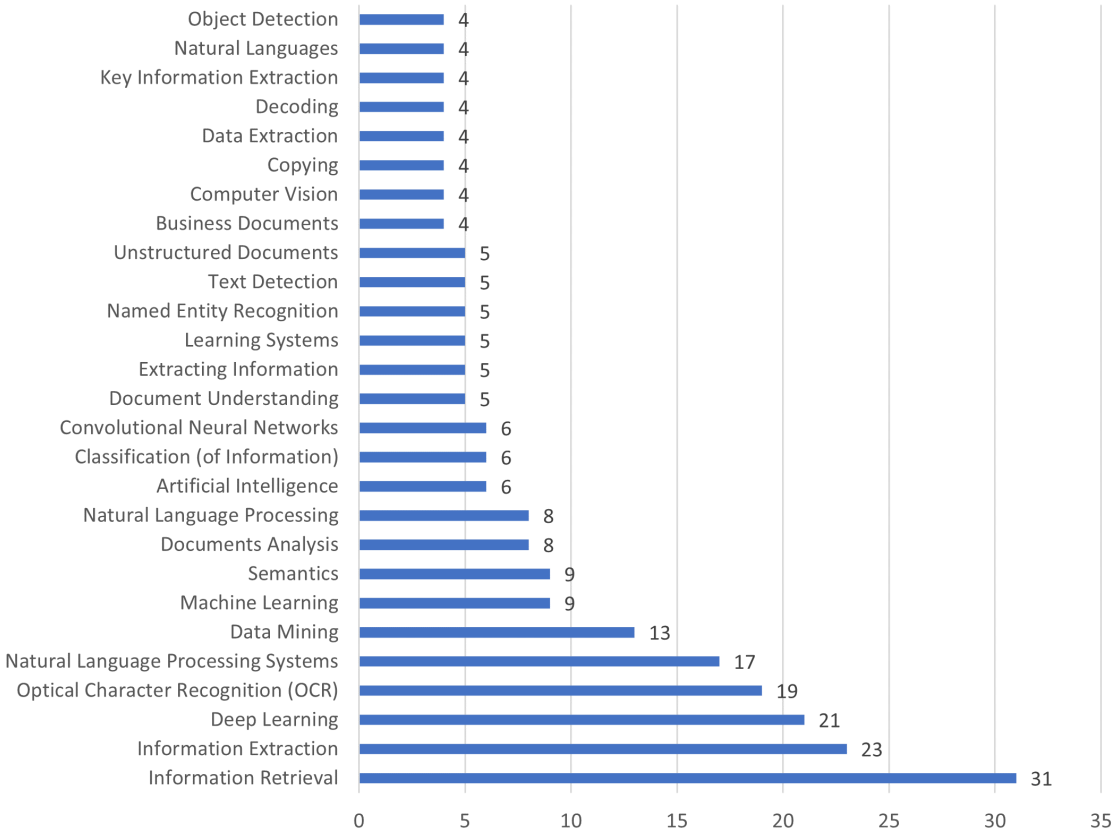


FIGURE 2.3. Documents by Keyword (Over 4 Publications).

techniques, such as rule-based or template-based approaches, rely on predefined rules or templates to identify and extract specific invoice information. However, these techniques do not have the ability to manage complex document formats in real-time scenarios, such as invoices and purchase orders [10]. On the other hand, ML approaches have revolutionized document management by enabling automatic learning of patterns from both textual content and visual elements. ML techniques are highly proficient in extracting valuable information from diverse document formats. With the help of sophisticated algorithms and models, we can analyze data and uncover meaningful insights that might otherwise go unnoticed. From textual analysis tasks like classification, Named Entity Recognition (NER), and sentiment analysis to visual pattern recognition tasks such as object detection, OCR, and layout analysis, ML facilitates the extraction of valuable insights.

In recent years, the use of ML techniques in the extraction of information from invoices has gained popularity due to its potential to automate and streamline the process. Although these systems do not require manual management, they still demand substantial amounts of data to achieve good outcomes [11]. Research in this field has shown that studies have suffered not only from small datasets but also from the quality of publicly available datasets. It is well documented, in the study by Krieger, Drews, and Funk [12], that machine learning models can be affected by layout bias and different models

are impacted to varying degrees by this bias. Pre-trained transformer models like LayoutLM [13] have outperformed other models in benchmarking for information extraction from invoices. This demonstrates their superiority in addressing these challenges.

As organizations strive to achieve greater efficiency and accuracy in handling invoices, researchers have searched for various techniques to extract essential information from these documents. Recent solutions involve end-to-end approaches that aim to complete a task without relying on multiple separate modules or stages.

This literature review aims to explore the use of ML techniques to extract information from invoices in different languages, optimizing the extraction process for scanned invoices, and addressing the challenges associated with extracting information from invoices with poor scan quality. Furthermore, the review will examine the differences in IE between scanned and digital invoices.

This section of the literature review is organized into two key categories to offer a thorough understanding of the tools and techniques utilized for extracting information from invoices. The first category, termed 1D tools, concentrates on text-based approaches that predominantly depend on the content and semantic structure of the invoice text. The second category, known as 2D tools, merges both text-based methods and layout-based features by factoring in the spatial arrangement of text and other elements within the invoice. These approaches capitalize on visual patterns and formatting cues—such as position, font size, and alignment—alongside the textual content to enhance the accuracy and reliability of information extraction. The upcoming subsections will delve into these two groups of tools, outlining their principles, strengths, and applications in relation to invoice data extraction.

2.3.1. One-dimensional Tools

One-dimensional tools make use of NLP to extract data from invoices. These tools typically rely on techniques such as NER, keyword matching, regular expressions, and semantic relationships to identify and extract relevant information from unstructured or semi-structured invoice documents. For instance, the NLP platform Appzen, described by Jain *et al.* [14], enables the processing and analysis of various transaction documents such as receipts, invoices, contracts, and purchase orders to detect unauthorized expenses. The techniques involved encompass ingestion, entity extraction, semantic analysis, audit, and feedback. Utilizing NLP alongside a lightweight semantic layer improves precision and effectiveness in identifying relationships and unauthorized expenses within financial transaction documents.

Despite these advancements, one-dimensional tools face substantial limitations in efficiently handling the complexity and variability of invoice formats. The diversity in language and structure of transaction documents often necessitates custom processing and analysis tailored to each type of document [14].

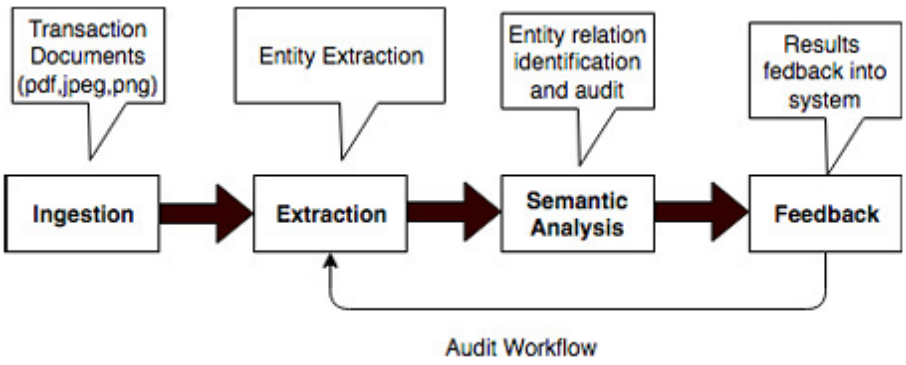


FIGURE 2.4. Four stages of audit workflow.

The workflow in Figure 2.4 illustrates the typical approach adopted by one-dimensional tools, emphasizing the sequential nature of the process and the reliance on NLP techniques for extracting and processing invoice data.

2.3.2. Two-dimensional Tools

Recent advancements in DL methods for document Key Information Extraction (KIE) tasks have led to the emergence of two major categories: grid-based approaches and graph-based approaches [15]. Grid-based approaches refer to a document representation method that is primarily based on an organized grid structure. This structure is characterized by horizontal and vertical lines that intersect to form cells, that include specific information. In contrast, Graph Neural Networks (GNN) Graph-based approaches use NLP techniques to represent text as nodes within a graph. Then these nodes are analyzed by a GNN to identify the relationships and connections between different entities. This approach allows for a more in-depth understanding of the document and can be particularly useful in tasks such as NLP, text classification, and document summarization. When analyzing document structure, it is often noted that invoices lack a standardized format. Invoices include details such as the date, items purchased, amounts, and Value Added Tax (VAT) taxes; however, this information may be located in different areas of the document based on each company's chosen format. As a result, IE tools will need to understand not only the relationship between the text but also the connections between different elements of the document [16]. Simpler approaches in this field employ object detection techniques to understand spatial information [17]. Studies in this domain suggest that object detection is a feasible method for IE from invoices. However, they acknowledge that future research could investigate the integration of NLP techniques, such as NER, to enhance the extraction of specific entities from invoice documents [17].

Grid-based

Katti *et al.* [18] approaches the issue from a different angle. Instead of converting a document into a linear text format, the proposed Chargrid method preserves the spatial layout of the document by encoding it as a sparse 2D grid of characters. This approach frames the

task of understanding documents as instance-level semantic segmentation on Chargrid. Specifically, the model is designed to predict a segmentation mask with pixel-level labels and object-bounding boxes to group multiple instances of the same class. The authors apply this Chargrid paradigm to an IE task from invoices and illustrate its effectiveness compared to state-of-the-art NLP algorithms and computer vision algorithms.

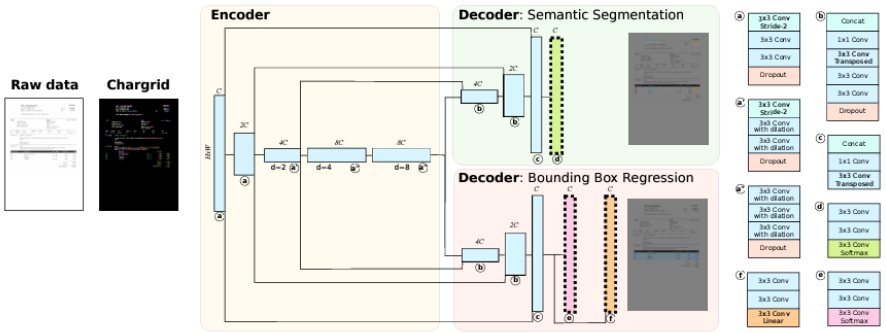


FIGURE 2.5. Workflow of the Chargrid method. The document is encoded as a 2D grid of characters, preserving spatial structure. The process includes semantic segmentation, pixel-level labeling, and object-bounding box predictions for information extraction.

Figure 2.5 illustrates the Chargrid workflow, highlighting how this method maintains the spatial layout of the document. By combining natural language processing and computer vision techniques, it offers a robust solution to extracting structured data from complex invoice formats.

In 2021, Bensch, Popa, and Spille [19] reconstructed the Chargrid [18] model to evaluated the potential advantages of using a box-bound regression decoder as an initial step to analyze the impact of individual components. Since BERTgrid [20] has demonstrated improved model performance through NLP pre-processing, this research also examined the benefits of NLP pre-processing on a generated dataset consisting of invoices. Therefore, the primary achievement of this work lies in offering a comparative analysis of cutting-edge models using a publicly accessible dataset. The study finds that NLP-based pre-processing improves model performance, but the use of a bounding box regression decoder only enhances performance for fields that do not follow a rectangular shape.

The study by Yeghiazaryan *et al.* [21] proposes a method called Tokengrid for more efficient data extraction from unstructured documents, focusing on invoices. The study begins by employing a grid-based approach to represent the document as a 2D grid. It then presents a method for the instance detection branch of the network, designed specifically for automatic IE from invoices. This simplifies the problem to 1D region detection, leading to a network with fewer parameters and shorter inference times.

When evaluating tasks that require word-level labels, the Attend, Copy, Parse architecture [22] has been shown to outperform a state-of-the-art production system based on word classification in the field of IE. The Attend, Copy, Parse architecture uses deep

neural networks and convolutional operations on concatenated document text and image modalities that can be trained directly on end-to-end data, bypassing the need for word-level labels. The architecture was evaluated on a large and diverse set of invoices, where it demonstrated superior performance compared to the existing production system based on distant supervision, word classification, and heuristics.

Graph-based

Detecting text, particularly in diverse, multi-language regions with indistinct characteristics, poses a significant challenge in computer vision and has gained attention recently. The article [23] presented a unified scheme for text localization and structured data extraction, integrating OCR and data mining techniques. The proposed Image Text Extraction (ITE) scheme combines DL and traditional data mining applications to detect text, predict word bounding boxes, and extract structured text data from various image layouts. This end-to-end trainable model effectively handles multi-scale and multi-lingual text, reducing manual labor and achieving high accuracy in real-world datasets.

Wei, He, and Zhang [24] mention in that digital documents, as opposed to scanned ones, preserve rich formatting information, offering significant advantages in terms of accessibility, usability, and data integrity. Unlike scanned documents, which essentially capture images of physical pages, digital documents retain the original electronic formatting, including fonts, styles, layouts, and metadata. However, it is important to note that the approach chosen by the authors in this article has limited exploration of layout semantics, thus not fully exploring spatial relationships between elements. The authors focus on developing a model that combines the power of large pre-trained language models and GNN to efficiently encode both textual and visual information in Visually Rich Documents (VRD).

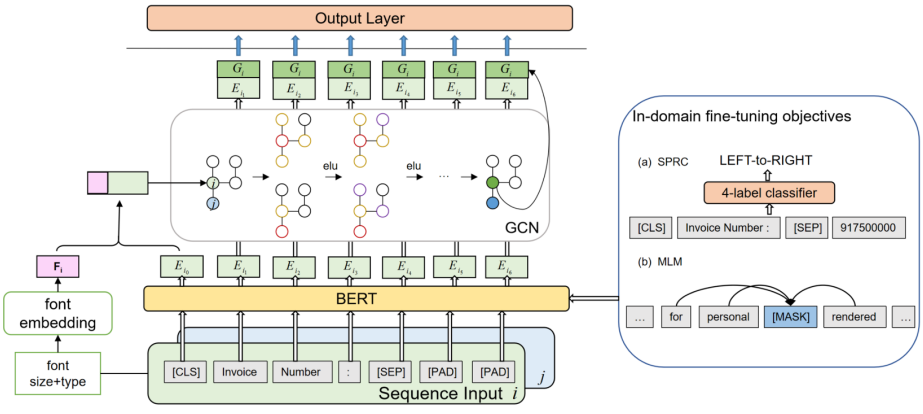


FIGURE 2.6. Workflow for the approach by [24] . The method integrates pre-trained language models with GNN to encode textual and visual features from digital documents while preserving rich formatting and metadata.

Figure 2.6 provides a visual representation of this workflow. The approach combines text encoding and visual data processing using a GNN, demonstrating how textual and

layout information can be effectively integrated for robust document understanding, even though layout semantics are not fully explored.

The top performing team at Research, Innovation and Vision for the Future (RIVF) 2021 Mobile-Captured Image Document Recognition (MC-OCR) for Vietnamese Receipts Competition, investigated the challenge of extracting structured information from low-quality scanned documents and its impact on IE system performance Nguyen, Nguyen, and Nguyen [25]. They developed a scorer to assess image quality and excluded the lowest quality invoices from training data, resulting in improvements in the designed model's performance.

Pham *et al.* [26] also utilized the same dataset from the MC-OCR Challenge 2021 to propose a solution using GNN for KIE from the Vietnamese invoices. This approach achieved results similar to those of the solutions presented in the competition.

Both Le, Pham, and Nguyen [27] and Paliwal *et al.* [28] present DL approaches to recognize scanned receipts. The goal is to automatically identify and extract key texts from the receipts to save them as structured data. Le, Pham, and Nguyen [27] system consists of two main modules: text detection based on Connectionist Text Proposal Network and text recognition based on Attention-based Encoder-Decoder. It also includes pre-processing to extract receipt areas and OCR verification to ignore handwriting. On the other hand, Paliwal *et al.* [28] TableNet is an end-to-end model that exploits the interdependence between table detection and structure recognition tasks.

Chua and Duffy [29] introduces Deep Conditional Probabilistic Context-Free Grammars (DeepCPCFG) for parsing complex two-dimensional documents. They employ Recursive Neural Networks to establish a comprehensive system that identifies the most likely parse representing the structured information to be extracted. DeepCPCFG is a method for end-to-end structured IE from business documents, such as invoices and receipts. It uses a Context-free Grammar (CFG) to model the information to be extracted, allowing it to capture complex recursive document structures. Unlike traditional approaches, DeepCPCFG does not rely on annotated bounding boxes, reducing the need for brittle post-processing and the re-engineering of the post-processing pipeline. The method yields compelling results and is effective in classifying bounding boxes, even when trained without any hand annotations. The effectiveness of DeepCPCFG suggests a need to re-evaluate grammars in image understanding, particularly when combined with DL

To better comprehend highly complex layouts Hwang *et al.* [30] paper presents an approach called SPAtial DEpendency parser (SPADE) to model spatial relationships and hierarchical information structure in semi-structured documents by leveraging linguistic and spatial information. This involves constructing a dependency graph with tokens and fields as graph nodes, which is then used to represent spatial dependencies. The methodology also incorporates the use of a Tail Collision Avoidance algorithm for improved performance.

Gemelli *et al.* [31] article presents Doc2Graph, which proposes a document layout analysis method based on graph representation. This framework constructs a fully connected graph over documents, with nodes representing words or semantic entities and edges showing pairwise relationships between them. Doc2Graph uses a language model to encode text and includes node and edge aggregation functions suitable for documents by leveraging the relative positioning of document objects through polar coordinates. These techniques allow the framework to extract important information from documents, like form understanding, invoice layout analysis, and table detection, without requiring extensive pre-training data.

In a different approach, Tata *et al.* [32] proposes a Scorer feature in his pipeline to score candidates for each field in the target schema, capturing the semantics of the fields. The scorer system in Glean is a ML model that takes as input the target field from the schema and the extraction candidate to produce a prediction score. On the basis of these learned candidate and field representations, each extraction candidate is scored based on the similarity to its corresponding field embedding.

Dang and Nguyen [33] also proposed a new DL architecture, Multi-Stage Attentional U-Net, for end-to-end IE on 2D character-grid embeddings. The model uses multi-stage encoder-decoders, self-attention mechanism, and box convolution to capture textual and spatial relations between 2D elements. A multi-task training scheme is also introduced to improve performance in challenging scenarios. The experimental results show that MSAU outperforms the baseline U-Net architecture using 40% fewer parameters. Additionally, extensive ablation studies verify the effectiveness of each component of the proposed model.

After examining various DL methods, Li, Jiang, and Song [34] determined that incorporating attention mechanisms into document feature analysis would enhance the precision and speed of document feature analysis. It was also observed that integrating and evaluating features from various sources and levels would enhance the effectiveness of document feature analysis.

The field of invoice IE is essential for automating business processes and improving efficiency, particularly in accounting. Despite the substantial amount of unstructured data available in businesses, its utility remains largely untapped. Research in invoice IE is broadly categorized into rule-based approaches, which use predefined patterns, and ML techniques, which enable the automated recognition of complex patterns in diverse document formats. While rule-based methods offer some precision, they struggle with varied invoice structures and real-time adaptation. Conversely, ML techniques, especially deep learning and neural networks, have transformed the field by providing more robust tools for tasks like NER, OCR, and layout analysis.

In recent years, the application of advanced ML models, including transformer-based architectures like LayoutLM [13], has gained traction due to their ability to manage the complexity and variability of invoice formats. However, challenges such as data scarcity,

layout biases, and inconsistent data quality still limit their effectiveness. To address these, researchers have explored hybrid solutions, combining rule-based and ML techniques, to capitalize on both approaches' strengths. There is also significant interest in one-dimensional tools focused on NLP techniques and two-dimensional tools that consider both text and spatial relationships within documents, using methods like grid-based and graph-based approaches.

State-of-the-art models in grid-based and graph-based document representation have made substantial progress in handling the spatial layout of documents. For example, grid-based models like Chargrid [18] and BERTgrid [20] offer structured 2D text analysis, while graph-based approaches employ GNN to capture spatial relationships, enhancing IE from complex invoice layouts. These advanced techniques, along with hybrid and end-to-end deep learning models, underscore the evolving landscape of invoice IE research, aiming to streamline extraction processes and improve the precision of document analysis across multilingual and multi-format datasets.

2.3.3. Large Language Models

LLMs, particularly those based on transformer architectures, have revolutionized the field of IE by enabling the analysis of complex documents with high accuracy. These models excel in tasks like NER, relationship extraction, and layout-aware information processing, making them ideal for semi-structured and unstructured documents such as invoices.

In the context of invoice IE, LLMs leverage their ability to understand both textual content and spatial relationships between elements in a document. Advanced models like LayoutLM [13] and LayoutLMv2 [35] incorporate multi-modal inputs, combining textual data with visual and layout information, which allows them to process a document as a whole rather than as isolated text. This approach is particularly effective in capturing intricate details in invoices, such as extracting key fields like invoice numbers, dates, and total amounts, even when they appear in varied formats or locations.

A comparative study by Douzon *et al.* [36] analyzed three prominent transformer-based models—LAMBERT [37], LayoutLM, and LayoutLMv2—focusing on their performance in extracting information from fiscal documents. LayoutLMv2 emerged as the top performer, achieving the highest f1-scores, precision, and recall, particularly excelling in complex datasets like the CORD dataset, where it attained an f1-score of 95.89%. However, its superior performance came at the cost of higher computational requirements, with a model size of 802.4 MB and an average inference time of 3.64 seconds. In contrast, LayoutLM, while less accurate, proved more computationally efficient and suitable for resource-constrained scenarios.

The study also emphasized the challenges posed by unstructured datasets, such as the Brazilian Receipt Dataset, which included handwritten text and irregular formats. LayoutLMv2 demonstrated resilience in handling these complexities, outperforming other models. The findings highlight the trade-offs between accuracy and efficiency, recommending LayoutLMv2 when accuracy is paramount and LayoutLM for environments with

limited resources. This underscores the importance of model selection based on specific use cases and computational constraints.

While transformer-based models like LayoutLMv2 showcase impressive capabilities in handling complex invoice layouts and integrating both textual and spatial information, they come with significant computational demands and rely heavily on multimodal data. This creates challenges in resource-constrained environments. Additionally, implementing and training such models often require substantial labeled data and specialized infrastructure, making them less feasible for certain applications.

In contrast, this study focuses exclusively on text-based approaches. It utilizes techniques such as regular expressions, keyword matching, and rule-based systems to tackle information extraction tasks. By concentrating on extracting insights directly from textual data, the proposed methodology aims to provide a lightweight, efficient, and practical solution tailored to the specific dataset provided by a private company. This approach emphasizes adaptability and simplicity, demonstrating the potential to achieve robust results without the complexities of multimodal architectures.

This thesis builds upon the foundations established by cutting-edge research, aiming to contribute to the field of invoice IE by tackling persistent challenges and investigating innovative approaches that cater to the intricacies of real-world data. The varied methodologies discussed in the literature, spanning from rule-based systems to advanced ML techniques, provide a comprehensive foundation for understanding the complexities involved in invoice processing. However, the limitations identified—such as difficulties in adapting to diverse invoice layouts, managing unstructured data, and addressing inconsistencies in data quality—highlight the urgent need for further innovation.

This work aspires to enhance invoice IE by prioritizing practical, scalable solutions that combine the strengths of both traditional and contemporary methods. Drawing inspiration from recent advancements in transformer-based models and hybrid techniques, the approach presented here focuses solely on text-based methodologies to ensure robust information extraction from the dataset provided by a private company. By examining and adapting strategies employed in grid-based and graph-based models, this thesis aims to bridge the divide between theoretical innovations and their real-world applicability.

Subsequent chapters outline the methods, tools, and the dataset utilized in this study, as well as the experiments conducted to validate the proposed approach. Particular emphasis is placed on addressing challenges such as misclassification of fields, and the extraction of key information, including totals, dates, and VAT numbers. By building on the strengths of existing models while addressing their limitations, this work contributes to the ongoing advancement of invoice IE research, focusing on achieving greater accuracy, flexibility, and efficiency in practical applications.

CHAPTER 3

Materials and Methods

The Materials and Methods chapter describes the approach and tools used to develop and test text-based machine learning techniques to extract key information fields from invoices: VAT number, invoice date, and total invoice value. This study used a combination of traditional regular expressions and LLMs to achieve accurate data extraction, highlighting the comparative strengths and limitations of each approach.

The dataset used for this work was provided by a private company and its proprietary nature introduced certain restrictions on data accessibility and privacy considerations. Therefore, the methodologies were designed not only to maximize extraction accuracy but also to adhere to these privacy restrictions, ensuring compliance with company data policies.

This chapter is organized to provide a systematic account of the pre-processing steps, algorithm selection, model training, and evaluation metrics applied. Each method, from regular expressions to machine learning algorithms, is explained in detail, offering insights into why specific techniques were chosen based on their alignment with the objectives of the project and the constraints of the data. The focus on reproducibility and data privacy serves as a foundation for validating the study findings and guiding future research in the field of automated document information extraction.

3.1. Data

The dataset used in this work was provided by a private company in the transport and logistics sector and comprises 1,354 PDF invoices from various industries and 96 different suppliers. Of these, 1,013 invoices are in Portuguese, while the remaining 341 are in other languages. The majority of the invoices are original PDFs; however, a subset consists of scanned invoices, which presented additional challenges during the extraction process. All invoices are written in the Roman alphabet, using only Latin letters, without the inclusion of special characters or alphabetic scripts from other writing systems. The dataset was divided into training and testing subsets using an 80/20 split, resulting in 809 Portuguese invoices and 273 invoices in other languages for training, with 204 and 68 invoices, respectively, allocated for testing.

Since January 1st, 2022, in Portugal, all invoices are required to have a Quick Response Code (QR Code). The QR Code for simplified invoices must follow the structure presented in table 3.1. Data from QR codes embedded in the Portuguese invoices was used to validate the extracted information, providing a reliable means of cross-checking the results. Consequently, only the Portuguese invoices were included in the final dataset

used for training and testing, as they are the only ones in the dataset that contain QR codes.

During preprocessing, several invoices were excluded due to specific challenges and system limitations. This included 164 invoices with unreadable formatting, 69 invoices that were excessively large and caused processing delays and system crashes, and 95 invoices with QR code reading errors that required manual intervention. These adjustments ensured that the dataset used was optimal for achieving reliable and efficient results in the information extraction process.

The system developed in this work focuses on extracting 3 fields:

(1) Field 'A': VAT of the issuer, must be filled in with the issuer's VAT, excluding any country prefix;

(2) Field 'F': Date of the document. The document date should be presented in the format YYYYMMDD;

(3) Field 'O': Total document amount with taxes should be formatted with two decimal places, using a period (".") as the decimal separator and without any thousands separator.

TABLE 3.1. Simplified Structure Defined by the Portuguese Tax Authority for Generating QR Codes.

Code	Description	Example	Mandatory
A	VAT of the issuer	A:123456789	+
B	VAT of the buyer	B:999999990	+
C	Country of the buyer	C:PT	+
D	Type of document	D:FS	+
E	Status of the document	E:N	+
F	Date of the document	F:20190812	+
G	Unique identification of the document	G:FS CDVF/12345	+
H	ATCUD	H:CDF7T5HD-12345	+
I1	Fiscal space	I1:PT	+
I7	Taxable base of VAT at the normal rate	I7:0.65	++
I8	Total VAT at the normal rate	I8:0.15	++
N	Total taxes	N:0.15	+
O	Total document amount with taxes	O:0.80	+
Q	4 characters of the Hash	Q:YhGV	+
R	Certificate number	R:9999	+
S	Other information	S:NU;0.80	++

In the context of information extraction from invoice texts, the complexity of extracting different fields—date, VAT number, and total value, varies significantly due to linguistic and formatting differences across invoices. These challenges are compounded by the multilingual and multi-currency nature of invoices, which affects the extraction accuracy for each field in distinct ways.

Date extraction presents fewer challenges compared to VAT numbers and total values. Despite the variations in date formats, such as YYYY-MM-DD, as shown in figure 3.1,

or other formats as DD/MM/YYYY or MM-DD-YYYY, the fundamental concept of a date remains relatively consistent across European countries and many other regions. European date formats, while varied, still adhere to recognizable patterns that can be detected using advanced regular expressions. For example, dates in invoices from Germany (DD.MM.YYYY) and France (DD/MM/YYYY) are similar in structure and can often be handled by a well-designed extration pattern. In contrast, date formats in the USA (MM-DD-YYYY) might differ, but the overall pattern recognition is feasible due to the regular nature of dates.

VAT number extraction presents a greater challenge due to significant variability in formats and structures across different countries. VAT numbers can differ not only in length but also in the format of the digits and letters used. For instance, in Portugal, VAT numbers (Número de Identificação Fiscal (NIF)) are typically nine digits long and start with the digit "5", such as 500 000 000. In contrast, in Germany for example, VAT numbers can include a country code followed by up to 11 digits (e.g., DE123456789). Similarly, in the UK, VAT numbers often begin with "GB" followed by a nine-digit number (e.g., GB123456789). In the USA, the equivalent is the Employer Identification Number (EIN), which can have a format such as 12-3456789. These variations require highly specialized patterns to accurately capture VAT numbers across different jurisdictions, complicating the extraction process.

The extraction of total values from invoices is also complex due to the wide range of formatting styles and the presence of multiple monetary values within the text. Total values can be expressed in various ways, such as with different currency symbols (e.g., €, $, £), separators (e.g., periods, commas), and formatting conventions. For instance, an invoice in France might present totals as "1 234,56 EUR," while a German invoice might list it as "1.234,56 €." Additionally, some invoices might include multiple totals or partial amounts, such as subtotals or additional fees or tax fees, which can further complicate the extraction process. The presence of these diverse formats and values requires a robust and adaptable approach to accurately capture the final total amount.

3.2. Methods

This study employs a structured processing pipeline designed to improve the accuracy and efficiency of information extraction from multilingual invoices for large-scale, multi-national applications. Initial evaluations revealed a scarcity of studies focused on one-dimensional tools capable of meeting the complex and multilingual demands of global enterprises. Recognizing these constraints, this work developed a multi-step, text-mining pipeline tailored to extract key fields—including VAT number, invoice date, and total invoice value—from invoices in diverse languages and formats.

The pipeline is divided into two primary stages: converting PDF invoices to text and extracting specific fields of knowledge. The PDF-to-text stage prepares raw invoice data for subsequent analysis, while the knowledge extraction stage applies a combination of

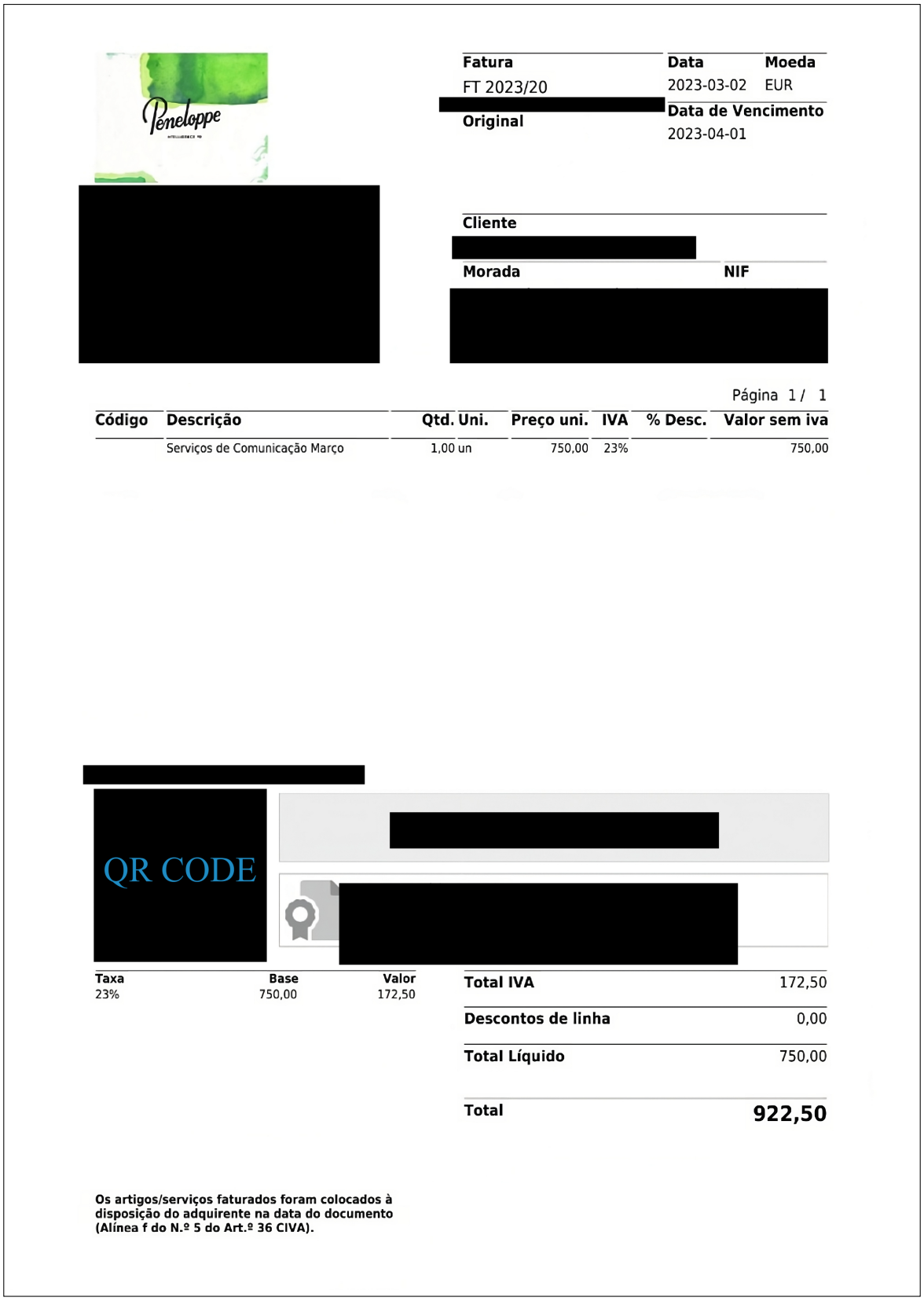


FIGURE 3.1. Invoice Example From the Portuguese Invoices Training Set.

rule-based and machine learning techniques to identify and retrieve relevant information accurately.

Figure 3.2 provides a visual representation of the process pipeline, illustrating each step in a clear and concise manner.

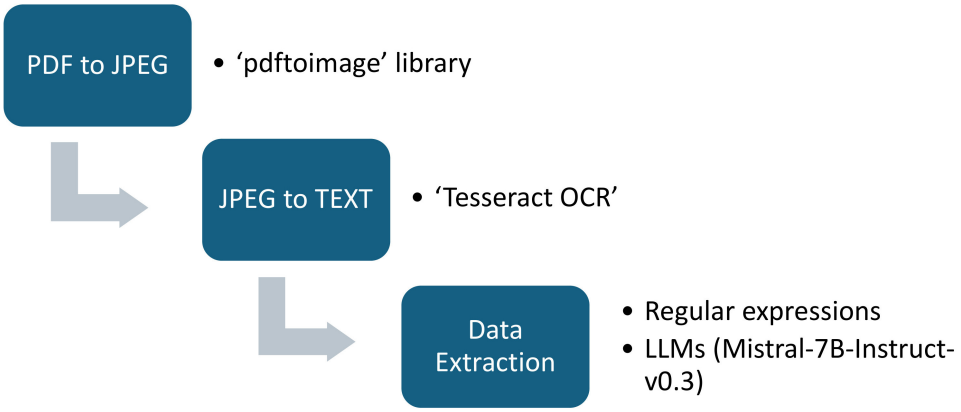


FIGURE 3.2. Processing Pipeline.

3.2.1. PDF to Text

The first stage in the processing pipeline converts Portable Document Format (PDF) files into textual data to enable further analysis. This step involves two key processes: transforming PDF into image format and then applying OCR to extract text.

The PDF files are initially converted to Joint Photographic Experts Group (JPEG) images using the pdf2image library. This Python-based, open-source library is particularly effective for applications requiring high-quality image rendering, such as OCR preprocessing. Initially, the conversion to Portable Network Graphics (PNG) format was tested; however, the subsequent steps of the pipeline encountered issues with a significant number of NA (Not Available) values, as the OCR process struggled to read these images effectively. As a result, it was decided to use JPEG images, which provided a more stable and reliable output for the OCR step. The pdf2image library was used in its default configuration, which is optimized for standard PDF to image conversion tasks, ensuring consistent results.

Following this, Tesseract OCR is applied to convert the images to text. Tesseract is a widely adopted open-source OCR engine that supports over 100 languages and is compatible with multiple image formats. The default version of Tesseract was employed for this task, utilizing the default Page Segmentation Mode (PSM) and the default language detection. PSM is a setting that defines how Tesseract interprets the layout of a page for text extraction. The default PSM (mode 3) assumes a single block of text and attempts to extract all text in the image without making assumptions about the layout. This configuration works well for documents with standard text flows, such as invoices, where the text is typically organized in a predictable manner. While more advanced PSM modes

could be used for documents with complex layouts, the default setting was chosen for its balance of speed and accuracy, given the structure of the invoices in this dataset.

Although it was not possible to test the veracity of the Tesseract output through manual validation of all extracted text, no errors were found in the extracted text during the training phase. For example, as shown in Figure 3.3, the original invoice image was correctly converted into text by Tesseract, and the output text in Figure 3.4 demonstrates that the OCR engine accurately captured the relevant information. The difficulties encountered in extracting the relevant information were not due to inaccuracies in the OCR result. Tesseract consistently extracted the text correctly, which allowed the focus to remain on other areas of the information extraction pipeline. This indicates that the OCR output itself was accurate and did not contribute to any misinterpretations of the data.

After completing the PDF-to-text conversion, the extracted text serves as the foundation for the knowledge extraction phase. This comprehensive approach ensures that textual data, regardless of its original document structure or language, is accurately captured and prepared for advanced analysis.

3.2.2. Knowledge Extraction

The knowledge extraction phase builds on the preprocessed text data to identify and extract specific fields of interest: VAT number, invoice date, and total invoice value. This phase progresses from basic rule-based techniques to advanced machine learning methods, each tailored to accommodate the unique characteristics of invoice data.

The initial approach in knowledge extraction utilizes regular expressions as a foundational tool for identifying predictable, structured patterns within the text. Regular expressions are particularly effective for recognizing fixed formats, such as dates and VAT numbers, where consistency in layout allows for precise extraction. This stage begins with straightforward regular expressions patterns designed to capture each target field individually, establishing a baseline performance in terms of accuracy and efficiency.

Following initial extraction, post-processing techniques are applied to refine the tokens extracted by regular expressions patterns. This iterative approach involves developing more sophisticated regular expressions patterns to account for slight variations in format, thereby enhancing the extraction's accuracy for fields with subtle inconsistencies.

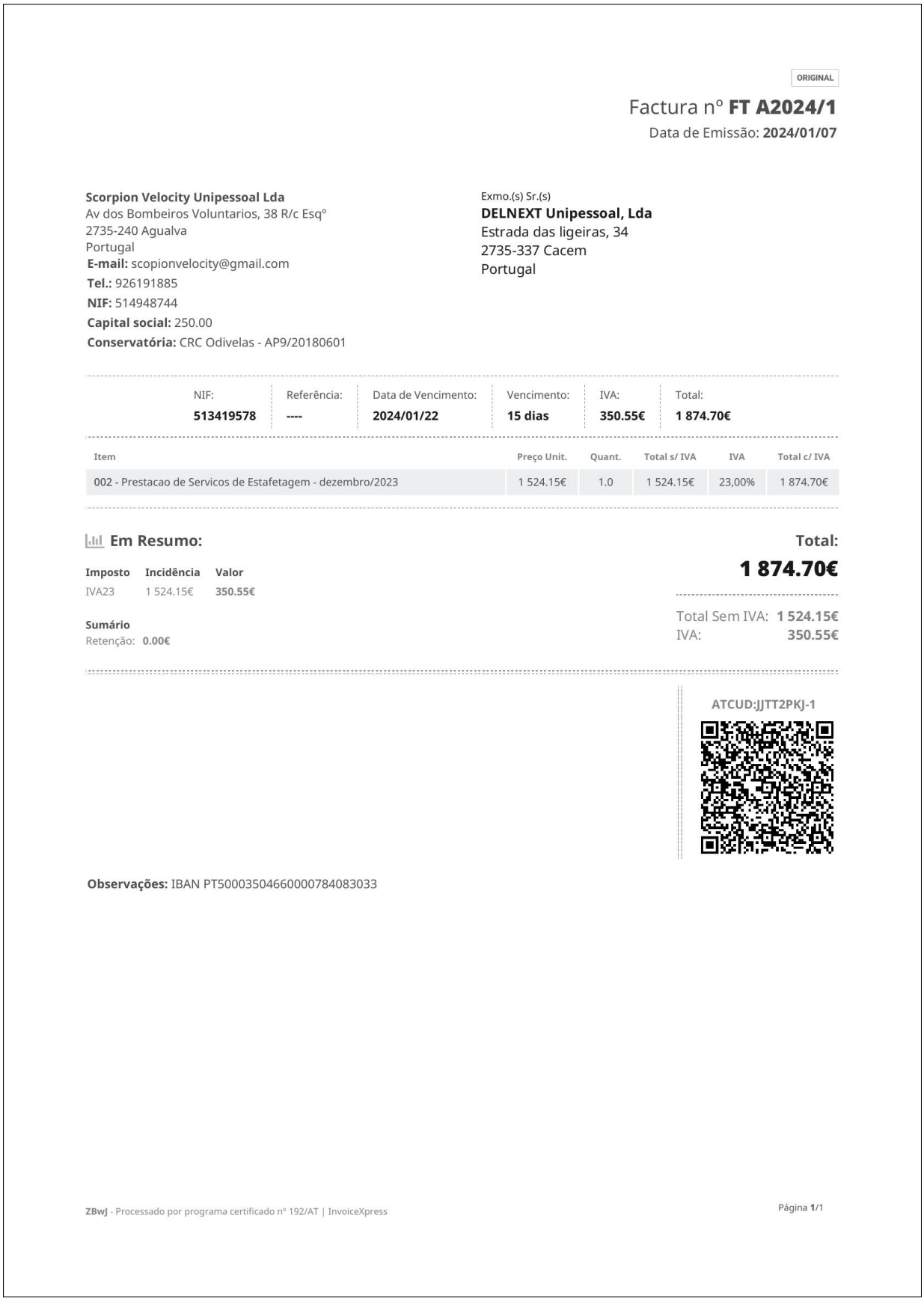
While regular expressions-based methods provide control and precision, they lack adaptability for unstructured or varied language data, necessitating more advanced techniques for the complex, diverse content typical of multinational invoices.

To address the limitations of rule-based extraction, the pipeline incorporates LLMs based on transformer architectures. LLMs are capable of understanding contextual nuances and semantic relationships within text, making them effective for handling multilingual and complex invoice data. The transition to LLMs allows for improved extraction accuracy by leveraging the models' ability to understand diverse linguistic structures beyond fixed patterns.

Generative models, known for their proficiency in generating and interpreting natural language, complement the LLMs by enabling more refined extraction of nuanced information that varies in format or context. This step represents a natural progression from simple pattern recognition to advanced machine learning, which enhances the robustness of the pipeline for unstructured and multilingual data.

By combining rule-based and machine learning approaches, this study explores the comparative strengths and limitations of each method, offering insights into the adaptability of text-mining techniques for diverse document formats. The pipeline thus provides a flexible and accurate system for extracting critical information from complex invoice data.

The primary metric for evaluating the effectiveness of the pipeline is accuracy, measured by the proportion of correctly identified items (VAT number, date, and total value) relative to the total number of items analyzed. Accuracy offers a straightforward and intuitive measure of performance, as each extraction attempt either succeeds or fails in identifying the required information. This focus on accuracy enables clear assessment of each technique's ability to meet the study's objectives in extracting critical fields from the invoice text.



ORIGINAL

Factura nº **FT A2024/1**

Data de Emissão: **2024/01/07**

Scorpion Velocity Unipessoal Lda
Av dos Bombeiros Voluntarios, 38 R/c Esqº
2735-240 Agualva
Portugal
E-mail: scopionvelocity@gmail.com
Tel.: 926191885
NIF: 514948744
Capital social: 250.00
Conservatória: CRC Odivelas - AP9/20180601

Exmo.(s) Sr.(s)
DELNEXT Unipessoal, Lda
Estrada das ligeiras, 34
2735-337 Cacem
Portugal

NIF:	Referência:	Data de Vencimento:	Vencimento:	IVA:	Total:
513419578	**----**	**2024/01/22**	**15 dias**	**350.55€**	**1 874.70€**

Item	Preço Unit.	Quant.	Total s/ IVA	IVA	Total c/ IVA
002 - Prestacao de Servicos de Estafetagem - dezembro/2023	1 524.15€	1.0	1 524.15€	23,00%	1 874.70€

Em Resumo:

Imposto	Incidência	Valor
IVA23	1 524.15€	**350.55€**

Sumário
Retenção: **0.00€**

Total:
1 874.70€

Total Sem IVA: **1 524.15€**
IVA: **350.55€**

ATCUD:JJTT2PKJ-1

Observações: IBAN PT50003504660000784083033

ZBwJ - Processado por programa certificado nº 192/AT | InvoiceXpress

Página **1**/1

FIGURE 3.3. Original invoice image

ORIGINAL

Factura n° FT A2024/1
Data de Emissdo: 2024/01/07

Scorpion Velocity Unipessoal Lda Exmo.(s) Sr.(s)

Av dos Bombeiros Voluntarios, 38 R/c Esq° DELNEXT Unipessoal, Lda
2735-240 Agualva Estrada das ligeiras, 34
Portugal 2735-337 Cacem

E-mail: scopionvelocity@gmail.com

Tel.: 926191885

NIF: 514948744

Capital social: 250.00

Conservat6ria: CRC Odivelas - AP9/20180601

Portugal

Data de Vencimento: : Vencimento: : :
2024/01/22 | 15 dias | 350.55€ | 1874.70€

NIF: i Referéncia:
513419578

Item PregoUnit. Quant. _—'Total s/IVA IVA Total c/ IVA
002 - Prestacao de Servicos de Estafetagem - dezembro/2023 1524.15€ 1.0 1 524.15€ 23,00% 1
874.70€
Em Resumo: Total:
Imposto Incidéncia_ Valor 1 874.70€
IVA23 1524.15€ SB0Sse rr ee rr ec rR eee ERR
a Total SemIVA: 1524.15€
Sumario .
Retencdo: 0.00€ IVA: 350.55€

ATCUD:JJTT2PKJ-1

Observacées: IBAN PT50003504660000784083033

ZBwJ - Processado por programa certificado n° 192/AT | InvoiceXpress Pagina 1/1

FIGURE 3.4. Text extracted from the invoice using Tesseract OCR

CHAPTER 4

Experiments and Results

This chapter details the experiments conducted to evaluate the effectiveness of the information extraction pipeline for invoices in multiple languages, organized in a systematic progression from simpler to more advanced methodologies. Each experiment was designed to assess the pipeline's ability to accurately and efficiently identify three key invoice fields: VAT number, invoice date, and total invoice value.

The experimental approach begins with regular expressions as a foundational tool, establishing a baseline for performance and providing precision and control over the extraction process in structured text environments. Regular expressions are highly effective in identifying patterns where the structure is consistent and predictable, making them a suitable starting point. In this initial phase, each field of interest was extracted using basic regex expressions, with postprocessing techniques applied to refine the extracted tokens further. As these experiments progressed, more complex regex patterns were implemented to capture nuanced variations in the data, enhancing extraction accuracy.

However, as the limitations of regular expressions became apparent, particularly its inability to adapt to variations in language or handle unstructured text, the experiments advanced to more sophisticated techniques, including LLMs. Transformer-based LLMs were applied to leverage contextual information and semantic relationships within the text, significantly improving extraction performance for diverse and complex data. Generative models, recognized for their capability to both generate and interpret natural language, were used to deepen the system's ability to understand and process intricate linguistic structures.

This progression from regular expressions to machine learning reflects a natural evolution in information extraction tasks, from rule-based methods suited to structured data to advanced machine learning techniques designed to handle unstructured multilingual content. By combining rule-based and ML methods, this study provides a comprehensive evaluation of the advantages and limitations of each approach, highlighting insights for the practical application of these techniques in real-world invoice processing scenarios.

To assess the effectiveness of each technique, accuracy was chosen as the primary metric. Accuracy provides an intuitive measure of performance by reflecting the proportion of correctly identified items out of the total analyzed, offering a straightforward assessment of each method's success in extracting the target fields. This metric is supplemented with additional considerations, such as processing efficiency and adaptability, to provide a balanced view of each method's strengths and potential applications.

4.1. Regular Expressions

In the field of information extraction, regular expressions play a crucial role in identifying and capturing specific patterns within textual data. Their strength lies in the ability to efficiently parse both structured and semi-structured text, making them particularly well-suited for initial extraction tasks where the data format is predictable. This section explores the application of regular expressions in the context of invoice processing, detailing the methodologies used to extract key fields such as VAT number, invoice date, and total invoice value.

The discussion begins with a thorough examination of the basic regular expression patterns employed for extraction, highlighting their precision and control in managing uniform data formats. Following this, more advanced regular expression techniques are introduced, demonstrating how refinements in pattern complexity enhance the ability to capture variations and exceptions within the text.

While regular expression methods are powerful tools for structured data, their inherent limitations require careful consideration of their role within the broader extraction pipeline. By establishing a solid foundation with regular expressions, this section sets the stage for transitioning to more sophisticated extraction techniques, thereby underscoring the evolution of methodologies employed throughout this study.

4.1.1. Baseline

In the initial phase of experimentation, this study concentrated on the application of straightforward regular expressions as a foundational method for extracting specific fields from textual data. The extraction process began with the development of basic regular expression patterns tailored to each field of interest, enabling the systematic retrieval of relevant information from the text.

Following the initial extraction, post-processing techniques were employed to enhance the quality of the extracted data. This refinement phase involved validating the accuracy of the identified patterns and making necessary adjustments to improve precision. By focusing on this baseline approach, the study established a critical reference point for evaluating subsequent extraction methods. This foundational work not only demonstrates the effectiveness of simple regular expressions but also sets the stage for more complex methodologies that will be explored in later phases of experimentation.

For the VAT number field, a regular expression was designed to identify variations of the VAT identifiers, disregarding case sensitivity and accommodating different formats such as 'NIF', 'VAT', or 'Contribuinte'. The associated function used these regular expressions to extract VAT numbers, excluding the VAT number of the buyer. The function returns the extracted VAT number or None if no valid VAT number was found.

In parallel, the date field was addressed using another regular expression, which targeted date formats commonly found in documents. A dedicated function applied this regular expression to identify and retrieve the oldest date among the first two matches in the text, offering flexibility and robustness in date extraction.

Furthermore, for the total value field, a third regular expression was employed to capture variations of total amounts within the text. A specialized function utilizing this regular expression was designed to identify and return the highest total value detected, accommodating different formats of numeric representation, whether separated by periods or commas.

The regular expression used can be found in Figure 4.1.

```
VAT_regex = r"(?i)(?:NIF|VAT|Contribuinte)[:\s]*(?:N\.?\s*|n.?°\s*)
    ?([0-9]{8,12})\b"
DATE_regex = r"\b\d{1,2}[-␣/]\d{1,2}[-␣/]\d{4}\b"
TOTAL_regex = r"(?i)Total.*?(\d{1,3}(?:\.\d{3})*(?:,\d{2})|\d{1,3}(?:,\d
    {3})*?:\.\d{2}))"
```

FIGURE 4.1. Baseline Regular Expressions for Each Field.

The evaluation of the regular expression mentioned in this section showed varying levels of accuracy across different fields, highlighting both its capabilities and limitations.

For the VAT number field, the method achieved an accuracy of 0.4760, indicating that the regular expression correctly identified the VAT number in approximately 47.60% of the cases. This performance suggests that, while the regular expression was able to capture some VAT numbers, it struggled with variations in format, reducing its overall effectiveness.

In contrast, the date field demonstrated a much lower accuracy of 0.1488, indicating significant difficulty in accurately extracting date information. In particular, of the cases where the regular expression did not identify a date, more than 83% were classified as Not a Number (NaN) values, indicating that the regular expression was largely unsuccessful in locating valid dates with the given pattern. This suggests that the primary challenge was not just the variety of date formats, but the inability of the regular expression to identify any dates within the given dataset, pointing to a need for more flexible or advanced methods to capture diverse date formats effectively.

The regular expression achieved an accuracy of 0.3612 for the Total Invoice value, correctly identifying the total values in 36.12% of cases. Among the cases where the regular expression did not return a match, 47% were classified as NaN values, while the remaining 53% were incorrect values. This distribution indicates that the regular expression encountered challenges not only in identifying total values but also in distinguishing them from other numeric data or formatting errors. Consequently, the performance of the regular expression underscores the challenges associated with accurately capturing total amounts, thereby emphasizing the necessity for more advanced techniques to enhance extraction reliability.

4.1.2. Advanced Pattern Extraction

As the study progresses, the focus shifts from basic regular expression techniques to advanced pattern extraction methods that address the inherent limitations of simpler

approaches. In this subsection, the exploration of sophisticated regular expressions will be detailed, emphasizing their ability to capture nuanced variations in the text that may arise from inconsistencies in formatting or language. This evolution in extraction methodology allows for a more robust and adaptable system capable of handling the complexities and diversities of real-world invoice data. By refining regular expression patterns to account for these variations, this phase aims to enhance the accuracy and reliability of the extracted information, ultimately contributing to the overall effectiveness of the information extraction pipeline.

Initially, the regular expression implemented for the VAT field was designed to capture VAT-like numbers by recognizing common prefixes such as "NIF," "VAT," or "Contribuinte," followed by a sequence of 8 to 12 digits. While this pattern proved effective for straightforward instances, it lacked the flexibility required to accommodate variations in which VAT numbers were presented with spaces or alternative formatting. Consequently, this limitation resulted in reduced accuracy, as the regular expression struggled to adapt to the diverse representations of VAT numbers found within the dataset.

To further enhance precision, an updated regular expression was specifically designed to capture VAT numbers that follow the Número de Identificação de Pessoa Coletiva (NIPC) format. The NIPC, assigned by the Registo Nacional de Pessoas Coletivas (RNPC), consists of nine digits, with the first digit ranging between 5 and 9, excluding the digit 7. This pattern is particularly relevant in the context of this study, as all Portuguese invoices analyzed are from private organizations, whose VAT numbers (NIPC) always begin with the digit "5." By focusing on this strict convention, the regular expression used ensures high precision in identifying VAT numbers specific to private organizations, accurately capturing cases that comply with the NIPC format. This targeted approach improves the extraction process by ensuring that only valid private organization-related VAT numbers are recognized.

The enhanced regular expression used to find the VAT numbers of the document is illustrated in Figure 4.2.

```
regex_VAT1 = r"(?i)(?:NIF|VAT|Contribuinte|cont\.|Número|NIPC)[:\s]*(?:N
    \.?\s*|n.?°\s*)?(?:pt)?(\d{1,3}(?:\s?\d{3})*\b)"

regex_VAT2 = r"(?<!\d)5\d{8}(?!\d)"
```

FIGURE 4.2. Complex Regular Expression for VAT Number Field.

The second round of tests with the enhanced VAT extraction methods demonstrated a marked improvement in accuracy. The accuracy for VAT extraction increased to 0.7535, reflecting a significant improvement over the initial approach. This level of accuracy represents the peak performance observed in the second round of testing, which highlights the efficacy of employing a synergistic approach with adaptable and specific patterns to address the various formats in which VAT numbers manifest. These findings emphasize the critical balance between versatility and precision in the application of regular expressions,

particularly regarding the enhancement of information extraction accuracy from invoice textual data, especially in semi-structured formats such as VAT numbers.

However, the extraction process for VAT numbers still encountered some challenges. A key issue was the resemblance between VAT numbers and other numerical fields on invoices, such as phone numbers. In Portugal, VAT numbers for private companies are always nine digits long, which can easily overlap with other fields that follow a similar numerical pattern. For instance, phone numbers, which are also nine digits long and often grouped in a format like "296 201 990", were occasionally misidentified as VAT numbers. This similarity increased the likelihood of false positives during the extraction process.

In addition to false positives, some cases returned NaN values due to instances where the VAT number was either not present in a recognizable format or surrounded by text that disrupted the regular expression's ability to isolate the correct field. These cases highlight the limitations of pattern-based extraction in handling edge cases or unconventional layouts.

For the date field, the initial regular expression approach demonstrated significant limitations in accurately identifying diverse date formats within the textual data of the invoices. The original method struggled particularly with the variety of date representations, such as different separators and mixed textual formats. To address these challenges, a more sophisticated regular expression pattern was developed to enhance the extraction process. This advanced regular expression was designed to handle a broader range of date formats and variations, recognizing that date information might not only appear in different numerical structures, but also include textual components that are not captured by simpler patterns. The objective was to create a more flexible and robust regular expression capable of accurately extracting dates from a diverse set of formats encountered in the dataset.

The enhanced regular expression for date extraction has been designed to encompass a broader range of date formats, substantially surpassing the functionality of the original regular expression. In addition to capturing a greater variety of separators compared to the initial pattern, the new regular expression also incorporates several numerical date formats, including:

- YYYY-MM-DD;
- DD-MM-YYYY.

In addition to these numerical formats, the enhanced regular expression also addresses cases where parts of the date might be expressed in textual form rather than strictly numerical format. This includes scenarios where months may be represented as words or abbreviations (e.g., "2023 Mar 15" or "15 Mar 2023"). By integrating these variations into the regular expression design, the approach aimed to significantly improve the accuracy of date extraction by accommodating the diverse ways dates can appear in natural language texts.

Figure 4.3 illustrates the regular expression used in this test.

```
r"""
\b\d{4}[-\/\.]\d{2}[-\/\.]\d{2}\b |
\b\d{2}[-\/\.]\d{2}[-\/\.]\d{4}\b |
\b\d{4}\s(?:jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec|
          jan|fev|mar|abr|mai|jun|jul|ago|set|out|nov|dez)\s\d{2}\b |
\b\d{2}\s(?:jan|feb|mar|apr|may|jun|jul|aug|sep|oct|nov|dec|
          jan|fev|mar|abr|mai|jun|jul|ago|set|out|nov|dez)\s\d{4}\b
"""
```

FIGURE 4.3. Complex Regular Expression for Date Field.

The application of the more complex regular expression for date extraction in the invoice text resulted in a substantial improvement, achieving an accuracy of 0.7318. This marks a significant increase compared to the initial approach, which had an accuracy of 0.1488. The enhanced regular expression ability to handle multiple date formats, such as YYYY-MM-DD, DD/MM/YYYY, and variations including different separators (dashes, slashes, dots and spaces), as well as cases where parts of the date are in text, contributed to this improved performance. By capturing a broader range of date representations and accounting for the diverse ways dates are expressed in the dataset, the regular expression demonstrated a much higher success rate in correctly identifying dates. This result underscores the importance of developing more flexible patterns in regular expression when dealing with semi-structured or unstructured text data, particularly in tasks like date extraction, where format variability is common.

Despite these advancements, extracting the correct date remains a challenging task. While 19% of the mismatches were due to NaN values where no recognizable date was identified, the majority of the errors stemmed from the incorrect selection of a date among the multiple dates often present in invoices. Commonly, invoices include dates for various purposes, such as issue date, due date, service date, budget date, or other related fields, all of which follow similar formats. This variability increases the likelihood of the regular expression misidentifying the intended date, particularly when contextual cues or surrounding text are not adequately accounted for.

Additionally, there were instances where errors in the OCR output contributed to the mismatches. Notably, some dates were erroneously recognized as belonging to the year 2028, a clear impossibility in this dataset, which likely stemmed from Tesseract misreading "2023" as "2028." Such errors underscore the limitations of relying solely on OCR output without incorporating additional validation mechanisms.

In the initial approach, the regular expression for extracting total values was designed to capture numerical amounts labeled "Total," focusing on commonly used currency formats. It targeted numbers with groups of three digits and allowed both periods (.) and commas (,) as separators, depending on whether they were used for decimal points or thousand-grouping. While effective for straightforward cases, this approach lacked flexibility, as it primarily focused on variations of the term "Total" and did not account for a broader range of labels or variations in whitespace and formatting.

To increase flexibility and accuracy, a more advanced approach was introduced in Figure 4.4, incorporating multiple patterns to capture a wider variety of formats.

```
regex_total_value1 = r"(?i)\bvalor\b[:\s]*[\r\n]*\s*(\d{1,3}(?:[.,\s]?\d
    {3})*\s*(?:[.,]\d{2})?)\s*¤?"

regex_total_value2 = r"(?i)(?:Total|EUR).*?(\d{1,3}(?:[.,\s]?\d{3})
    *(?:[.,]\d{2})?)\s*¤?"

regex_total_value3 = r"(?i)(?:\btotal\b.*?|\bvalor\b.*?|\beur\b.*?)(\d
    +[.,]\d{2})\s*(?:¤|EUR)?"

regex_valores_soltos = r"(?m)^\d{1,3}(?:[.,\s]?\d{3})?(?:[.,\s]?\d{2})¤?$"
```

FIGURE 4.4. Complex Regular Expression for Total Value Field.

The first pattern expanded beyond "Total" to include variations of the word "valor" (value), accommodating different amounts of whitespace and line breaks, which improved detection in less structured or more complex text environments. The second pattern further broadened the scope by targeting keywords such as "Total" and "EUR" while allowing for various separators, such as spaces, periods, or commas, making the expression adaptable to different formatting conventions.

The third pattern increased versatility by recognizing numerical values associated with keywords like "total", "valor", or "EUR", while accounting for minor variations in punctuation. Lastly, an additional pattern was introduced to capture standalone numerical values not explicitly labeled, addressing cases where the total amount appeared without a specific descriptor.

The employed function subsequently returned the maximum value among all matches detected by these regular expression patterns. This method guaranteed the extraction of the most relevant total value, despite multiple potential matches.

This more comprehensive approach significantly improved the extraction process, allowing for a broader range of descriptors, formatting styles, and contextual variations compared to the initial, more rigid method. By using a combination of these patterns, the extraction method became more flexible and adaptable, enabling more accurate identification of total values from a diverse set of textual data.

The refined extraction method achieved an accuracy of 0.6636 for the total value, reflecting a significant improvement over the initial approach. This result demonstrates the effectiveness of the comprehensive pattern recognition strategy in capturing a wider array of total values from various formatting styles and textual contexts. However, despite this progress, the total value proved to be the most challenging field to accurately extract, not because of the presence of NaN values but because the method still returned incorrect values in a substantial number of cases. This indicates that further refinement is needed to improve the accuracy and reliability of total value extraction, highlighting the ongoing need for enhanced techniques to address the complexities of this particular field.

Among the errors identified, 47% of the mismatches were attributed to NaN values, where the regular expressions failed to identify a valid total value in the text. This highlights the limitations of the current patterns, which, despite being advanced, still do not account for the full variability in invoice structures and formats. These issues often arise from variations in the presentation of total values, such as unconventional labels, inconsistent formatting, or the proximity of irrelevant text elements that obscure direct identification.

The remaining 53% of the mismatches involved incorrect values being extracted. In some instances, the method mistakenly identified numerical patterns resembling the total value but were unrelated. Common examples include the extraction of VAT numbers, International Bank Account Number (IBAN), or share capital values, which are often longer numerical sequences. These can inadvertently match the patterns due to their similar structure and prominent placement on the invoice. Additionally, there were cases where the method returned values lower than the actual total amount, as it matched intermediary numbers, such as quantities, subtotals, or totals without VAT, which often appear near the desired total value in the invoice text.

These challenges underscore the inherent complexity of accurately extracting the total value field using purely text-based approaches. The presence of multiple numerical fields, varying layouts, and overlapping patterns increases the likelihood of false positives and missed matches.

4.2. LLMs and Generative Models: Contextual and Semantic Understanding

In the search for better results and more flexible solutions for information extraction from invoice text, the study transitioned from traditional rule-based methods like regular expressions to more advanced models such as LLMs. While regular expression provided a precise and deterministic approach for extracting specific patterns from structured text, it struggled with the variability and complexity inherent in real-world invoices, which can differ significantly in format, structure, and language. The limitations of regular expression, particularly in handling unstructured or semi-structured text with ambiguous content, drove the need for more sophisticated techniques. LLMs, with their ability to understand and generalize across diverse inputs, emerged as a strong alternative for extracting relevant data from these complex documents.

For this study, which focuses on information extraction from invoice text, we selected the Mistral-7B-Instruct-v0.3 model for several key reasons. Unlike cloud-based models like Bidirectional Encoder Representations from Transformers (BERT) or Generative Pre-trained Transformer (GPT)-3, Mistral can be deployed and executed locally, offering enhanced privacy and security—essential for protecting sensitive financial data contained within the datset used. The model's ability to operate locally ensures that no data needs to be transferred to external servers, a critical requirement for handling the confidential information. Additionally, Mistral's fine-tuning for instruction-following tasks makes it especially effective for extracting specific pieces of information, such as invoice numbers,

dates, and payment terms, from varied and unstructured text formats, making it a superior choice for this study's needs over other models.

The Mistral-7B-Instruct-v0.3 is an instruct fine-tuned version of the base model Mistral-7B-v0.3, a Large Language Model (LLM) with 7 billion parameters. Fine-tuning the model using instruction-based learning aligns its capabilities to better handle tasks involving user prompts in natural language, making it well-suited for applications where the model needs to follow directives or perform specific tasks based on human instructions.

This model inherits the architecture of Mistral-7B, which is based on a transformer model architecture similar to GPT-like models. It is designed for tasks like language modeling, text generation, and natural language understanding. The instruct fine-tuning process involves training the base model on instruction-based datasets to improve alignment with user instructions and generalization across diverse tasks that can include question answering, summarization, data extraction, classification, and other NLP tasks.

Mistral-7B-Instruct-v0.3 is a transformer-based language model that uses self-attention mechanisms to capture contextual information. During fine-tuning for instruct tasks, the model learns to infer task-related patterns by being exposed to pairs of inputs and target outputs in the form of directives and expected results, improving its ability to follow human-given instructions in various contexts. This is a medium-sized model optimized for understanding and responding to instructions. It is adaptable and performs well in various natural language processing tasks such as text generation, summarization, question answering, classification, extraction, and code generation. Additionally, the model is pre-trained on diverse datasets to generalize across multiple domains.

The model is particularly well-suited for use cases that involve direct interaction between a user and the model, where users expect precise, task-oriented responses. Its instruction-following capabilities are optimized for, for example, virtual assistants, automated customer support, educational tools, content creation, and data extraction and summarization.

Mistral-7B-Instruct-v0.3 is a highly capable model, but it does have some inherent limitations. Firstly, while the 7B parameters strike a good balance between capability and efficiency, it may not perform as well as larger models on more complex, nuanced tasks. Additionally, the quality of the model's output is highly dependent on the clarity and specificity of the instructions provided. Ambiguous or poorly phrased prompts can lead to inferior responses. Lastly, Mistral-7B-Instruct-v0.3 can still generate biased or incorrect information, especially when responding to tasks that require factual accuracy or involve sensitive topics.

In the initial experiment with the Mistral-7B-Instruct-v0.3 model, a simple prompt was used to establish a baseline for the model's performance in extracting key information from invoices. The prompt focused on extracting three specific fields from a provided invoice text: the invoice date (formatted as YYYY-MM-DD), the total value of the invoice, and the VAT number of the acquirer. The objective of the prompt was to be clear and direct,

testing the model's capacity to comprehend and extract specific details without any extra information or instructions. The exact prompt used is shown in Figure 4.5.

```
prompt = (
    "This is an invoice.\n"
    "What is the invoice text?\n"
    "Invoice Text: {invoice_text}\n"
    "Please extract the following information:\n"
    "* Invoice Date (YYYY-MM-DD format)\n"
    "* Invoice Total Value\n"
    "* VAT number of the acquirer"
)
```

FIGURE 4.5. Baseline Prompt.

Following the initial output from the model, a post-processing phase was implemented to refine and standardize the extracted data. This step was necessary because the raw output from the model often contained variations in format, which could affect the accuracy of subsequent analyses. Three main functions were developed for post-processing: 'clean_currency', 'clean_vatnumber', and 'convert_date'.

The 'clean_currency' function aimed to extract the numeric part of the total value output, as the model might return values with currency symbols or extraneous characters. Using regular expressions, the function isolated numeric patterns and converted them into floating-point numbers for consistency. This ensured that the extracted values could be reliably used in further calculations or comparisons.

The 'clean_vatnumber' function was applied to the VAT numbers extracted by the model. Since VAT numbers often include non-numeric characters or formatting inconsistencies, this function removed any leading non-digit characters and attempted to convert the remaining string into a numeric format. This helped standardize the VAT numbers, making them easier to compare against expected values. However, if the conversion failed, the original extracted value was retained to avoid data loss.

For dates, the 'convert_date' function was employed to transform date strings into a standardized format. The function parsed dates in the 'YYYY-MM-DD' format and converted them into a numeric representation (YYYYMMDD.0). This approach facilitated comparison across different dates, allowing for more accurate evaluation of the model's performance in extracting and interpreting date information.

The results of this baseline experiment provided valuable insights into the model's initial capabilities. The accuracy for extracting VAT numbers was relatively low, at 38.29%, indicating that the model struggled with identifying and correctly formatting VAT numbers from the provided invoice text. The accuracy of date extraction, however, was significantly higher at 75.97%, suggesting that the model was more capable of recognizing and parsing date information from invoices. The extraction of the total invoice value achieved an accuracy of 63.26%, demonstrating moderate success in identifying and interpreting numerical values associated with the total amount.

Overall, this initial experiment highlighted areas where the model performed well and aspects that required further refinement. The relatively low accuracy in VAT number extraction suggests that the model may benefit from additional training or prompts that provide more specific guidance on identifying such identifiers. Meanwhile, the stronger performance in date extraction offers a promising foundation for further development. These results serve as a baseline, establishing a reference point for future improvements in both prompt design and post-processing techniques.

In the second test conducted with the Mistral-7B-Instruct-v0.3 model, a more refined prompt was utilized to enhance the accuracy of information extraction from invoices. This prompt was designed to provide additional context and specific instructions on how to format the output. The template used for this experiment is as demonstrated in Figure 4.6.

```
prompt2 = (
    "This is an invoice.\n"
    "Invoice Text: {invoice_text}\n"
    "Please extract the following information and format your response
    as shown below:\n"
    "Invoice Date: YYYY-MM-DD (return the oldest date if multiple are
    found)\n"
    "Invoice Total Value: VALUE (return the highest value found)\n"
    "VAT Number of the Acquirer: VAT_NUMBER (8-12 digits, may include
    country code, remove spaces and non-numeric characters, must be
    different from '513419578')"
)
```

FIGURE 4.6. Detailed Prompt.

This more detailed prompt not only instructed the model to extract specific pieces of information but also clarified how to handle multiple entries. For example, it requested that if multiple dates were present, the model should return the oldest date, and if multiple total values were found, it should return the highest one. Furthermore, the prompt provided guidelines for formatting the VAT number, emphasizing the need to exclude non-numeric characters and ensure it differed from a specified example.

Following the model's output, post-processing was applied to further refine the extracted information. The functions employed in this phase included the 'convert_date' function, which was similar to that used in the first test, ensuring that dates were parsed and formatted consistently. This function attempted to convert date strings in the 'YYYY-MM-DD' format into a numeric representation (YYYYMMDD.0), facilitating easier comparisons.

In addition to the date conversion, several data normalization steps were implemented using pandas for improved accuracy and consistency in the results. For VAT numbers, the extracted values were converted to numeric types using 'pd.to_numeric()', which ensured that any non-numeric characters were handled appropriately, with errors coerced

into NaN values. A similar approach was taken for the total invoice values, where non-numeric entries were also converted to NaN, enabling a cleaner dataset for analysis.

For date handling, the extracted dates were converted into datetime objects to facilitate proper comparison, and the time component was stripped to focus solely on the date. This normalization was crucial for accurately determining the oldest date among potential multiple entries.

The results of this second test revealed a mixed performance across the three target fields compared to the initial test. Specifically, the VAT number field extraction accuracy improved to 44.81%, which, while still modest, represented an increase from the 38.29% observed in the first test. However, the accuracy for invoice date field extraction decreased to 65.12%, down from 75.97% in the previous experiment. This decline suggests that the model faced challenges with the increased complexity of the prompt and the specific requirements for handling multiple dates. In contrast, the accuracy for the total value field extraction improved significantly to 71.01%, compared to the 63.26% recorded in the first test. This enhancement indicates that the refined prompt's emphasis on extracting the highest value contributed positively to performance in this area.

While the improvements in the VAT number and the total value fields extraction are encouraging, the decline in the date field extraction accuracy highlights the need for further refinement in the model's approach to processing more complex prompts. This continuous process of refining prompts and enhancing post-processing techniques is crucial for creating an efficient automated tool for invoice processing.

4.3. Discussion

This work explored the efficacy of various text-based information extraction methods for invoices in multiple languages, addressing the challenges posed by diverse formats and languages. We investigated two main methods: traditional regular expressions and pre-trained language models, each with distinct approaches to invoice data extraction. The following section delves into the results, comparing each method's strengths and weaknesses in extracting the three selected fields: VAT numbers, dates, and total values.

Table 4.1 summarizes the accuracy of each method for extracting VAT numbers, dates, and total values, allowing a comparative analysis of their performance across fields.

TABLE 4.1. Extraction Accuracy of VAT Number, Date, and Total Value Fields Using Different Techniques.

Extraction Technique/Invoice Field	VAT Number	Date	Total Value
Regular Expressions			
Baseline	47.60%	14.88%	36.12%
Advanced Pattern	75.35%	73.18%	66.36%
LLMs and Generative Models			
Simple Prompt	38.29%	75.97%	63.26%
Refined Prompt	44.81%	65.12%	71.01%

The initial approach applied simple regular expressions, resulting in a VAT accuracy of 47.60%, date accuracy of 14.88%, and total value accuracy of 36.12%. These relatively low accuracies, particularly for dates, indicate that regular expressions, while capable of identifying straightforward patterns, struggle to generalize across the diverse formatting of multilingual invoices.

In contrast, advanced regular expressions improved the results significantly, achieving a VAT accuracy of 75.35%, date accuracy of 73.18%, and total value accuracy of 66.36%. This marked improvement highlights that sophisticated regular expressions when carefully crafted, can capture more complex data patterns. These findings demonstrate the effectiveness of advanced regular expressions for structured fields, where consistent and tailored pattern matching yields higher accuracy.

The second phase of the study focused on evaluating the Mistral-7B-Instruct-v0.3 model with two prompting strategies. The initial simple prompt achieved a VAT accuracy of 38.29%, date accuracy of 75.97%, and total value accuracy of 63.26%, revealing the model's ability to extract dates while underperforming in VAT extraction.

By introducing contextual cues in the prompt, VAT accuracy improved to 44.81% and total value accuracy to 71.01%, though this adjustment led to a decline in date accuracy to 65.12%. These results suggest that while contextual prompts can enhance specific fields, they may compromise other areas, especially when the model is balancing multiple extraction tasks. This implies that pre-trained language models are sensitive to prompt design, requiring precise tailoring to maintain consistent performance across varied extraction fields.

Each field achieved the highest accuracy with a different method, highlighting the unique strengths of each approach. Advanced regular expressions outperformed the language model in structured fields such as VAT numbers and, to a lesser extent, total values, demonstrating their utility in predictable, highly-structured data extraction. By contrast, the language model, particularly with prompt adjustments, proved more effective in extracting date information, where format variability benefits from the model's natural language processing capabilities. These results suggest that while advanced regular expressions excel in fields with consistent structures, language models may offer advantages in fields requiring more linguistic comprehension and flexibility.

The findings from this comparative analysis underscore the importance of selecting extraction methods based on the specific data field characteristics. Regular expressions are well-suited to fields like VAT numbers that benefit from predictable patterns, while language models, with prompt fine-tuning, are effective for fields like dates that require interpretive understanding of contextual cues. Together, these insights suggest the potential for hybrid approaches that could integrate the strengths of both techniques to enhance overall extraction accuracy across diverse invoice fields.

CHAPTER 5

Conclusions and Future Work

In today's digital economy, the vast amount of data generated daily has become both a resource and a challenge for businesses. In invoice processing, unstructured data extraction can streamline accounting and financial workflows, reduce manual labor, and improve data accuracy, but this task remains highly complex due to the variability in formats and languages. The primary aim of this study has been to evaluate automated methods to support the reliable extraction of essential data fields from invoices in multiple languages, facilitating more efficient financial and operational processes within organizations.

The demand for automated solutions in invoice processing has become especially pressing in the accounting sector, where the manual extraction of information from invoices is a time-consuming and error-prone task. Accounting professionals are required to carefully examine each invoice to obtain key details such as invoice number, date, and total amount, making the process costly and challenging to scale for businesses handling large volumes. Automation through machine learning promises significant improvements by reducing manual intervention, thereby enhancing accuracy and freeing up resources for more strategic tasks [4]. Additionally, as many organizations operate globally, they face unique challenges with multi-language invoices that traditional extraction methods often struggle to address.

5.1. Conclusions

This study recognizes the strategic benefits of unstructured data processing tools for applications such as real-time audits, fraud prevention, and forecasting accuracy [2]. By investigating the methods for invoice information extraction, this work aims to support more efficient financial data management, thereby enabling more informed decision-making and optimized workflow efficiency.

The motivation for this research stems from the substantial operational benefits that organizations could realize through the automation of invoice data extraction. Traditional methods, which rely on manual processing, are prone to error, time-consuming, and resource-intensive, often resulting in delayed or inaccurate financial reporting. Automation offers an opportunity to improve these workflows, enhancing the precision of financial data and reducing processing times, thereby supporting more agile financial operations. However, despite advances in multi-dimensional models combining text and structural data, there remains a gap in one-dimensional, text-only extraction tools. These simpler models could provide accessible and efficient alternatives where full document structure analysis is unnecessary or resources are limited.

This study investigated methods for automating information extraction from multilingual invoices, with a specific focus on extracting VAT numbers, dates, and total values. To address this challenge, it focused on developing a one-dimensional, text-based extraction tool aimed at filling a gap in existing research, with the objective of creating a scalable solution capable of catering to diverse organizational needs, including those of smaller businesses. By limiting the investigation to text-based techniques, this work aimed to deliver a resource-efficient option that would improve data accessibility and support accurate decision-making across organizations of various sizes and technical capabilities.

Guided by these goals, the study sought to answer two primary research questions:

(1) How can text-based information extraction methods be optimized to accurately extract relevant fields (e.g., invoice number, date, total amount) from multilingual invoices?

The results indicate that advanced regular expressions provide a robust solution for fields with predictable structures, such as VAT numbers and, to some extent, total values. These methods achieved the highest accuracy in VAT extraction due to the ability of carefully crafted patterns to handle consistent alphanumeric formats across multilingual invoices. Advanced regular expressions also demonstrated improved performance over simpler patterns, capturing nuances like formatting variations and extraneous information, thereby enhancing extraction reliability.

For fields with more variability, like total values and dates, language models—particularly with refined prompts—showed promise. While language models did not outperform regular expressions in VAT extraction, they performed comparably or even superiorly in extracting dates and total values, especially when contextual information was included in the prompts. These findings suggest that the optimization of text-based extraction methods depends significantly on the characteristics of the target data field: regular expressions are preferable for structured data with consistent patterns, while language models excel in fields where format variability requires a more flexible understanding.

(2) How effective are pre-trained multilingual models for information extraction tasks on invoice text data?

Pre-trained language models, like the Mistral-7B-Instruct-v0.3, demonstrated moderate effectiveness for certain fields, achieving the highest accuracy in date extraction. This indicates that language models can be effective for tasks requiring interpretive flexibility, as they can recognize and adapt to varied date formats and language-specific differences. However, the language model's effectiveness was sensitive to prompt design; while a refined prompt increased accuracy for some fields (notably total values), it decreased accuracy for others (dates), underscoring the model's reliance on carefully tailored prompts. Although pre-trained language models can process language with flexibility, their effectiveness

in structured field extraction, like VAT, was limited when compared to advanced regular expressions.

The results of this study, while contributing valuable insights into text-based information extraction methods, indicated lower accuracies compared to existing advanced one-dimensional extraction models. For instance, research conducted by Pham *et al.* [26] achieved extraction accuracies of 93% for timestamps and 87.4% for total values, highlighting the potential for enhancing text-only methods through the integration of image-based techniques. This suggests that while the development of a one-dimensional extraction tool offers a practical solution for many organizations, there remains significant room for improvement, particularly when considering the performance benchmarks set by more sophisticated models.

5.2. Limitations and Future Work

These discrepancies emphasize the need to recognize the limitations of this work. One significant limitation was the restricted diversity of the dataset used for the investigation. The dataset was sourced from a private company, which meant that it consisted solely of invoices from their specific suppliers within a specific industry sector. As a result, the formatting and language variations present in the invoices were limited to those encountered by this particular company. This lack of diversity in the dataset may have constrained the study's ability to comprehensively evaluate the performance of the extraction methods across a broader spectrum of invoice formats and languages. Consequently, the findings may not be fully representative of the challenges associated with invoice data extraction in a more diverse context, where organizations often face a wider range of invoice layouts and linguistic variations.

Another important limitation arose from privacy considerations associated with the private company. The necessity to maintain confidentiality restricted the use of cloud-based LLMs. This requirement led to the selection of LLMs that could operate locally, thereby limiting the range of models that could be employed. The larger and more advanced models, which typically offer enhanced performance due to their extensive parameter sizes, were not feasible for deployment in this context. As a result, the research may not have leveraged the full potential of state-of-the-art LLMs, potentially impacting the accuracy and effectiveness of the extraction methods evaluated in this study.

These limitations highlight the need for future research to explore diverse datasets and advanced LLMs in various environments to validate and enhance the findings presented here. Addressing these factors will be crucial in developing more robust and adaptable solutions for automated invoice processing across a broader range of organizational contexts.

Future research could further enhance the findings of this study by addressing current limitations, particularly by expanding the dataset to include a broader range of invoice formats, layouts, and languages. Such diversity would enable models to adapt

more effectively across diverse real-world scenarios, increasing robustness and accuracy in information extraction tasks.

Addressing privacy concerns is another crucial area for future work. Given the sensitive nature of financial data, compliance with privacy regulations is paramount, and currently restricts the use of larger, more advanced models. Future studies could explore techniques such as data anonymization, federated learning, or secure multi-party computation to enable the use of sophisticated models while safeguarding data privacy. Implementing these approaches would allow for more complex model architectures without compromising confidentiality, thereby broadening the scope of viable extraction methods.

Additionally, fine-tuning pre-trained language models on invoice-specific data presents another promising avenue. Fine-tuning could enable these models to achieve a higher degree of accuracy and contextual understanding, particularly in identifying nuanced information like invoice dates and total values. This customization may lead to improvements over generic language models by adapting to the unique structure and terminology found in financial documents.

The exploration of hybrid models that integrate both rule-based and machine-learning approaches holds significant potential for improving information extraction. A hybrid approach could dynamically select the most suitable method based on the characteristics of the field being extracted. For example, structured fields with predictable formats, such as VAT numbers, could benefit from the precision and efficiency of advanced regular expressions. In contrast, fields requiring contextual interpretation, such as dates and total amounts, could leverage the adaptability of fine-tuned language models. By intelligently combining these techniques, such a system could harness the strengths of both approaches, creating a more versatile and accurate extraction pipeline tailored to the specific requirements of each field.

In conclusion, this study contributes valuable insights into the application of regular expressions and language models for multilingual invoice data extraction. The findings underscore the need for method selection based on data characteristics and indicate that future research into dataset expansion, privacy-preserving methodologies, model fine-tuning, and hybrid approaches could significantly optimize information extraction accuracy. Continued advancements in these topics will support organizations in achieving greater operational efficiency and accuracy in financial data processing.

References

[1] B. Gärtner and M. R. W. Hiebl, "Issues with big data," in *The Routledge Companion to Accounting Information Systems*, M. Quinn and E. Strauss, Eds., 1st ed., Routledge, Dec. 22, 2017, pp. 161–172, ISBN: 978-1-315-64721-0. DOI: 10.4324/9781315647210-13. [Online]. Available: https://www.taylorfrancis.com/books/9781317297345/chapters/10.4324/9781315647210-13 (visited on 03/10/2024).

[2] A. Faccia, L. P. L. Cavaliere, P. Petratos, and N. R. Mosteanu, "Unstructured over structured, big data analytics and applications in accounting and management," in *Proceedings of the 2022 6th International Conference on Cloud and Big Data Computing*, Birmingham United Kingdom: ACM, Aug. 18, 2022, pp. 37–41, ISBN: 978-1-4503-9657-8. DOI: 10.1145/3555962.3555969. [Online]. Available: https://dl.acm.org/doi/10.1145/3555962.3555969 (visited on 03/10/2024).

[3] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," *Journal of Business Research*, vol. 69, no. 2, pp. 897–904, Feb. 2016, ISSN: 01482963. DOI: 10.1016/j.jbusres.2015.07.001. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0148296315002842 (visited on 03/10/2024).

[4] School of Business, Clark Atlanta University, USA, S. K. Herath, D. Woods, and School of Business, Clark Atlanta University, USA, "Impacts of big data on accounting," *The Business and Management Review*, vol. 12, no. 2, Dec. 25, 2021, ISSN: 20472862, 20472854. DOI: 10.24052/BMR/V12NU02/ART-15. [Online]. Available: https://cberuk.com/cdn/conference_proceedings/2022-01-20-18-24-55-PM.pdf (visited on 03/10/2024).

[5] D. Baviskar, S. Ahirrao, V. Potdar, and K. Kotecha, "Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions," *IEEE Access*, vol. 9, pp. 72 894–72 936, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3072900. [Online]. Available: https://ieeexplore.ieee.org/document/9402739/ (visited on 02/03/2024).

[6] T. Saout, F. Lardeux, and F. Saubion, "An overview of data extraction from invoices," *IEEE Access*, vol. 12, pp. 19 872–19 886, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3360528. [Online]. Available: https://ieeexplore.ieee.org/document/10418145/ (visited on 03/10/2024).

[7] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson,

S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting, and D. Moher, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, n71, Mar. 29, 2021, ISSN: 1756-1833. DOI: 10.1136/bmj.n71. [Online]. Available: https://www.bmj.com/lookup/doi/10.1136/bmj.n71 (visited on 01/21/2024).

[8] C. Parent. "Council post: Why businesses already have all the data they need," Forbes. Section: Innovation. (), [Online]. Available: https://www.forbes.com/sites/forbestechcouncil/2023/12/13/why-businesses-already-have-all-the-data-they-need/ (visited on 01/27/2024).

[9] "Analytics and AI-driven enterprises thrive in the age of with," Deloitte Insights. (), [Online]. Available: https://www2.deloitte.com/us/en/insights/topics/analytics/insight-driven-organization.html (visited on 01/27/2024).

[10] D. Baviskar, S. Ahirrao, and K. Kotecha, "Multi-layout invoice document dataset (MIDD): A dataset for named entity recognition," *Data*, vol. 6, no. 7, p. 78, Jul. 20, 2021, ISSN: 2306-5729. DOI: 10.3390/data6070078. [Online]. Available: https://www.mdpi.com/2306-5729/6/7/78 (visited on 01/21/2024).

[11] A. Hamdi, E. Carel, A. Joseph, M. Coustaty, and A. Doucet, "Information extraction from invoices," in *Document Analysis and Recognition – ICDAR 2021*, J. Lladós, D. Lopresti, and S. Uchida, Eds., Series Title: Lecture Notes in Computer Science, vol. 12822, Cham: Springer International Publishing, 2021, pp. 699–714, ISBN: 978-3-030-86330-2. DOI: 10.1007/978-3-030-86331-9_45. [Online]. Available: https://link.springer.com/10.1007/978-3-030-86331-9_45 (visited on 02/03/2024).

[12] F. Krieger, P. Drews, and B. Funk, "Automated invoice processing: Machine learning-based information extraction for long tail suppliers," *Intelligent Systems with Applications*, vol. 20, p. 200 285, Nov. 2023, ISSN: 26673053. DOI: 10.1016/j.iswa.2023.200285. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2667305323001102 (visited on 01/21/2024).

[13] Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou, "LayoutLM: Pre-training of text and layout for document image understanding," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Virtual Event CA USA: ACM, Aug. 23, 2020, pp. 1192–1200, ISBN: 978-1-4503-7998-4. DOI: 10.1145/3394486.3403172. [Online]. Available: https://dl.acm.org/doi/10.1145/3394486.3403172 (visited on 02/03/2024).

[14] P. Jain, K. Verma, A. Gaikwad, and P. Gadde, "Understanding financial transaction documents using natural language processing," in *Proceedings of the 10th International Conference on Knowledge Capture*, Marina Del Rey CA USA: ACM, Sep. 23, 2019, pp. 255–258, ISBN: 978-1-4503-7008-0. DOI: 10.1145/3360901.3364439. [Online]. Available: https://dl.acm.org/doi/10.1145/3360901.3364439 (visited on 01/21/2024).

[15] L.-W. Thiée, F. Krieger, and B. Funk, "Extraction of information from invoices – challenges in the extraction pipeline,"

[16] S. Bhat, D. Adiga, M. Shah, and V. Vyeth, "Effectiveness of visual features on diverse reading orders for information extraction," in *2019 IEEE 5th International Conference on Computer and Communications (ICCC)*, Chengdu, China: IEEE, Dec. 2019, pp. 1759–1764, ISBN: 978-1-72814-743-7. DOI: 10.1109/ICCC47050.2019.9064294. [Online]. Available: https://ieeexplore.ieee.org/document/9064294/ (visited on 01/21/2024).

[17] A.-S. Bulzan and C. Cernazanu-Glavan, "Object detection in invoices," in *2022 26th International Conference on System Theory, Control and Computing (ICSTCC)*, Sinaia, Romania: IEEE, Oct. 19, 2022, pp. 656–662, ISBN: 978-1-66546-746-9. DOI: 10.1109/ICSTCC55426.2022.9931900. [Online]. Available: https://ieeexplore.ieee.org/document/9931900/ (visited on 01/21/2024).

[18] A. R. Katti, C. Reisswig, C. Guder, S. Brarda, S. Bickel, J. Höhne, and J. B. Faddoul, "Chargrid: Towards understanding 2d documents," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 4459–4469. DOI: 10.18653/v1/D18-1476. [Online]. Available: http://aclweb.org/anthology/D18-1476 (visited on 01/21/2024).

[19] O. Bensch, M. Popa, and C. Spille, "Key information extraction from documents: Evaluation and generator," arXiv, Jun. 9, 2021. arXiv: 2106.14624[cs]. [Online]. Available: http://arxiv.org/abs/2106.14624 (visited on 01/21/2024).

[20] T. I. Denk and C. Reisswig, "BERTgrid: Contextualized embedding for 2d document representation and understanding," Sep. 2019.

[21] A. Yeghiazaryan, K. Khechoyan, G. Nalbandyan, and S. Muradyan, "Tokengrid: Toward more efficient data extraction from unstructured documents," *IEEE Access*, vol. 10, pp. 39 261–39 268, 2022, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2022.3164674. [Online]. Available: https://ieeexplore.ieee.org/document/9749071/ (visited on 01/21/2024).

[22] R. B. Palm, F. Laws, and O. Winther, *Attend, copy, parse – end-to-end information extraction from documents*, in arXiv, Apr. 23, 2021. arXiv: 1812.07248[cs]. [Online]. Available: http://arxiv.org/abs/1812.07248 (visited on 01/21/2024).

[23] Y. Ye, S. Zhu, J. Wang, Q. Du, Y. Yang, D. Tu, L. Wang, and J. Luo, "A unified scheme of text localization and structured data extraction for joint OCR and data mining," in *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, WA, USA: IEEE, Dec. 2018, pp. 2373–2382, ISBN: 978-1-5386-5035-6. DOI: 10.1109/BigData.2018.8622129. [Online]. Available: https://ieeexplore.ieee.org/document/8622129/ (visited on 01/21/2024).

[24] M. Wei, Y. He, and Q. Zhang, "Robust layout-aware IE for visually rich documents with pre-trained language models," arXiv, May 22, 2020. arXiv: 2005.11017[cs]. [Online]. Available: http://arxiv.org/abs/2005.11017 (visited on 01/21/2024).

[25] D.-C. Nguyen, T.-A. Nguyen, and X.-C. Nguyen, "MC-OCR challenge 2021: End-to-end system to extract key information from vietnamese receipts," in *2021 RIVF International Conference on Computing and Communication Technologies (RIVF)*, Hanoi, Vietnam: IEEE, Aug. 19, 2021, pp. 1–5, ISBN: 978-1-66540-435-8. DOI: 10.1109/RIVF51545.2021.9642083. [Online]. Available: https://ieeexplore.ieee.org/document/9642083/ (visited on 01/21/2024).

[26] V. D. Pham, L. Q. Nguyen, N. T. Pham, B. H. Nguyen, D. N. Minh Dang, and S. D. Nguyen, "Key information extraction from mobile-captured vietnamese receipt images using graph neural networks approach," in *2022 6th International Conference on Green Technology and Sustainable Development (GTSD)*, Nha Trang City, Vietnam: IEEE, Jul. 29, 2022, pp. 232–237, ISBN: 978-1-66546-628-8. DOI: 10.1109/GTSD54989.2022.9989111. [Online]. Available: https://ieeexplore.ieee.org/document/9989111/ (visited on 01/21/2024).

[27] A. D. Le, D. V. Pham, and T. A. Nguyen, "Deep learning approach for receipt recognition," in *Future Data and Security Engineering*, T. K. Dang, J. Küng, M. Takizawa, and S. H. Bui, Eds., Series Title: Lecture Notes in Computer Science, vol. 11814, Cham: Springer International Publishing, 2019, pp. 705–712, ISBN: 978-3-030-35652-1. DOI: 10.1007/978-3-030-35653-8_50. [Online]. Available: http://link.springer.com/10.1007/978-3-030-35653-8_50 (visited on 01/21/2024).

[28] S. Paliwal, V. D, R. Rahul, M. Sharma, and L. Vig, "TableNet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images," arXiv, Jan. 6, 2020. arXiv: 2001.01469[cs,eess]. [Online]. Available: http://arxiv.org/abs/2001.01469 (visited on 01/21/2024).

[29] F. C. Chua and N. P. Duffy, "DeepCPCFG: Deep learning and context free grammars for end-to-end information extraction," arXiv, Jun. 5, 2021. arXiv: 2103.05908[cs]. [Online]. Available: http://arxiv.org/abs/2103.05908 (visited on 01/21/2024).

[30] W. Hwang, J. Yim, S. Park, S. Yang, and M. Seo, "Spatial dependency parsing for semi-structured document information extraction," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, 2021, pp. 330–343. DOI: 10.18653/v1/2021.findings-acl.28. [Online]. Available: https://aclanthology.org/2021.findings-acl.28 (visited on 01/21/2024).

[31] A. Gemelli, S. Biswas, E. Civitelli, J. Lladós, and S. Marinai, "Doc2graph: A task agnostic document understanding framework based on graph neural networks," vol. 13804, 2023, pp. 329–344. DOI: 10.1007/978-3-031-25069-9_22. arXiv: 2208.11168[cs]. [Online]. Available: http://arxiv.org/abs/2208.11168 (visited on 01/21/2024).

[32] S. Tata, N. Potti, J. B. Wendt, L. B. Costa, M. Najork, and B. Gunel, "Glean: Structured extractions from templatic documents," *Proceedings of the VLDB Endowment*, vol. 14, no. 6, pp. 997–1005, Feb. 2021, ISSN: 2150-8097. DOI: 10.14778/3447689.3447703. [Online]. Available: https://dl.acm.org/doi/10.14778/3447689.3447703 (visited on 01/21/2024).

[33] T.-A. N. Dang and D.-T. Nguyen, "End-to-end information extraction by character-level embedding and multi-stage attentional u-net," arXiv, Sep. 9, 2021. arXiv: 2106.00952[cs,eess]. [Online]. Available: http://arxiv.org/abs/2106.00952 (visited on 01/21/2024).

[34] Y. Li, W. Jiang, and S. Song, "Review of semi-structured document information extraction techniques based on deep learning," presented at the Proceedings - 2023 2nd International Conference on Machine Learning, Cloud Computing, and Intelligent Mining, MLCCIM 2023, 2023, pp. 112–119. DOI: 10.1109/MLCCIM60412.2023.00022. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85182023401&doi=10.1109%2fMLCCIM60412.2023.00022&partnerID=40&md5=4e0e0a62515e62e8fc493715ee048bc6.

[35] Y. Xu, Y. Xu, T. Lv, L. Cui, F. Wei, G. Wang, Y. Lu, D. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou, *LayoutLMv2: Multi-modal pre-training for visually-rich document understanding*, Jan. 10, 2022. DOI: 10.48550/arXiv.2012.14740. arXiv: 2012.14740[cs]. [Online]. Available: http://arxiv.org/abs/2012.14740 (visited on 12/23/2024).

[36] T. Douzon, S. Duffner, C. Garcia, and J. Espinas, "Improving information extraction on business documents with specific pre-training tasks," vol. 13237, 2022, pp. 111–125. DOI: 10.1007/978-3-031-06555-2_8. arXiv: 2309.05429[cs]. [Online]. Available: http://arxiv.org/abs/2309.05429 (visited on 01/21/2024).

[37] Ł. Garncarek, R. Powalski, T. Stanisławek, B. Topolski, P. Halama, M. Turski, and F. Graliński, "LAMBERT: Layout-aware (language) modeling for information extraction," in vol. 12821, 2021, pp. 532–547. DOI: 10.1007/978-3-030-86549-8_34. arXiv: 2002.08087[cs]. [Online]. Available: http://arxiv.org/abs/2002.08087 (visited on 12/23/2024).